

PSYCHOLOGICAL

STATISTICS

QUINN MCNEMAR



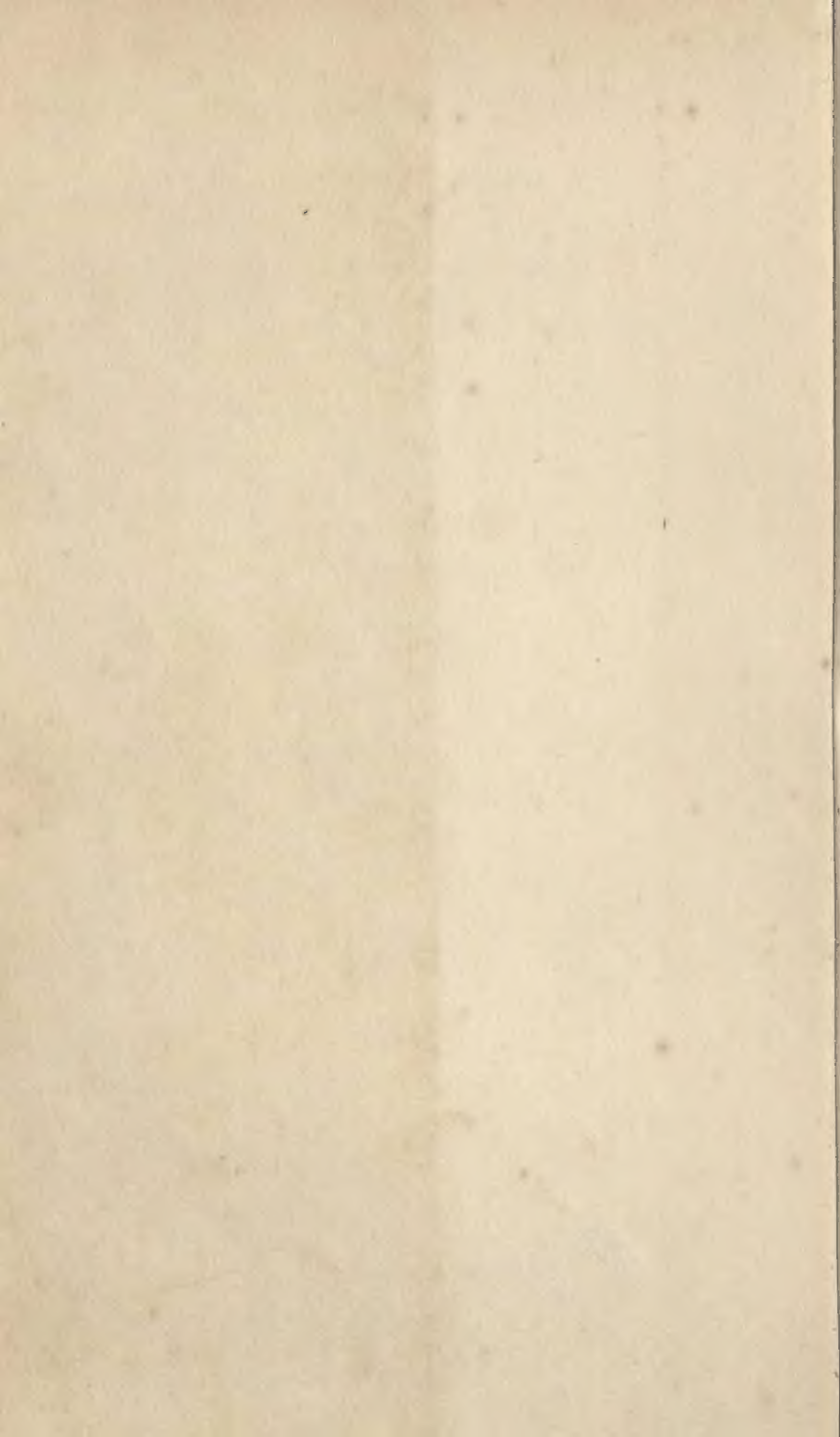


805

311  
MCN

A286  
mlb9





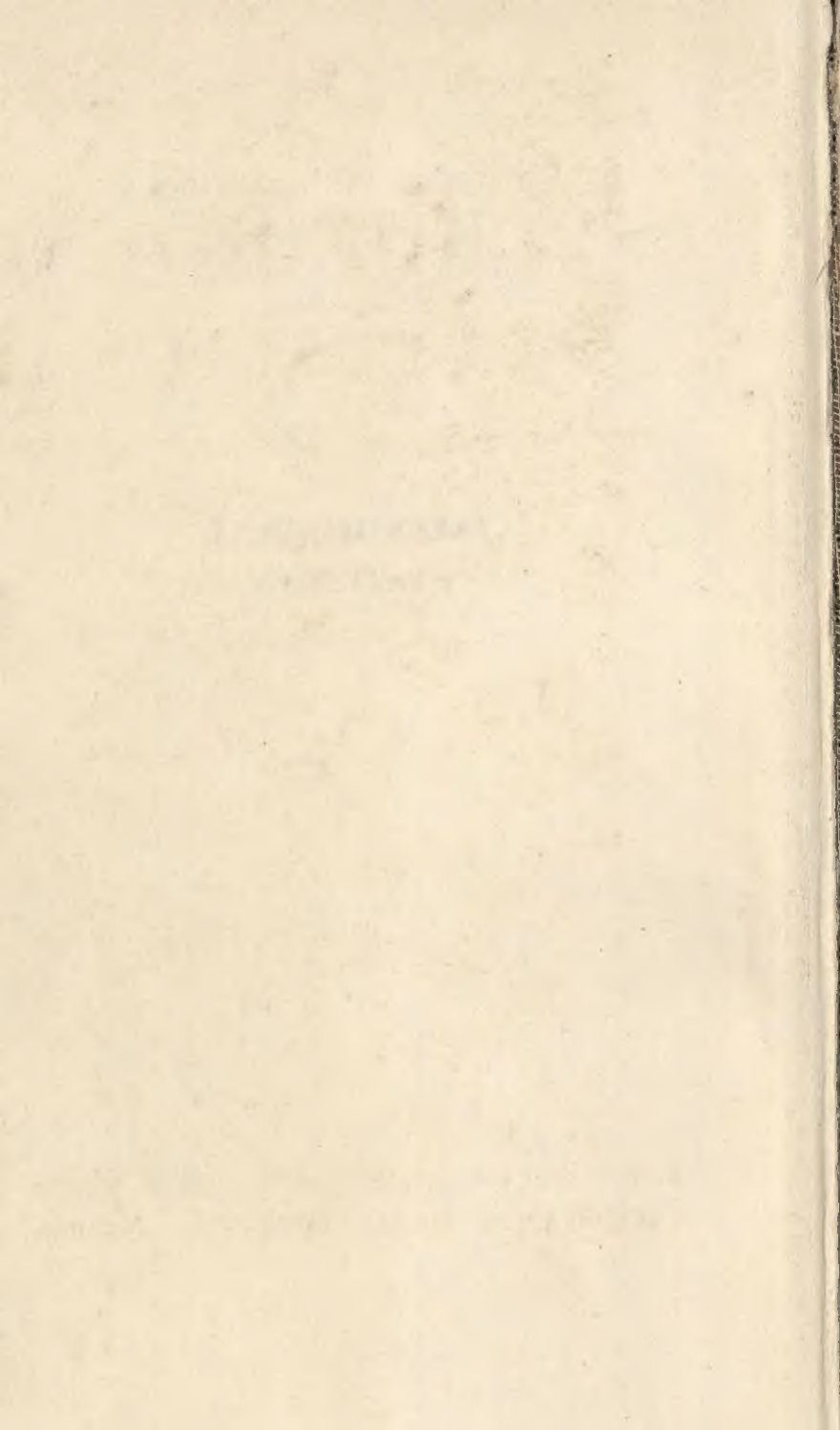


A WILEY PUBLICATION  
IN PSYCHOLOGY

HERBERT S. LANGFELD

*Advisory Editor*

**PSYCHOLOGICAL  
STATISTICS**



# PSYCHOLOGICAL STATISTICS

---

Second Edition

---

QUINN MCNEMAR

PROFESSOR OF PSYCHOLOGY,  
STATISTICS, AND EDUCATION  
STANFORD UNIVERSITY

JOHN WILEY & SONS, INC., NEW YORK  
CHAPMAN & HALL, LIMITED, LONDON

200.402  
11-21  
C 85

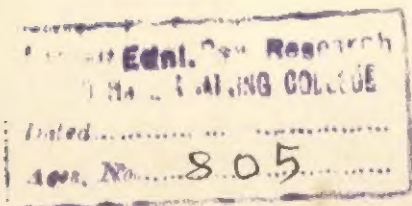


311  
MCN

COPYRIGHT, 1949, 1955,  
BY  
JOHN WILEY & SONS, INC.

*All Rights Reserved*

*This book or any part thereof must not  
be reproduced in any form without  
the written permission of the publisher.*



Library of Congress Catalog Card Number: 55-6320

PRINTED IN THE UNITED STATES OF AMERICA

## Preface

The widespread adoption of the first edition of this textbook suggests that it has been found useful in introductory courses although it was not written primarily for that level. In this revised and enlarged edition the elementary treatment of statistical inference has been expanded so as to make the book more palatable to the beginning student. An attempt has been made in Chapter 5 to introduce the student to the general problems associated with hypothesis testing. These concepts are further developed and extended to continuous variables for the large-sample situation in Chapter 6 and for small samples in Chapter 7. It is my firm conviction that it is better pedagogically to provide a separate treatment of small-sample techniques, and in this order.

In the development of the logic of statistical inference in Chapter 5, more use has been made of the binomial distribution, and some of the Neyman-Pearson principles of hypothesis testing have been introduced, with the notion of point estimation and confidence intervals postponed to Chapter 6.

The five chapters devoted to correlational analysis contain numerous minor revisions and extensions. The previous introduction to chi square has been replaced by a binomial approach, and a method for handling several correlated proportions and the exact probability method for fourfold tables have been added. A short chapter presents methods for comparing both correlated and independent variabilities, including Bartlett's test for homogeneity of variance and the  $F$  distribution.

The first chapter on the analysis of variance is essentially unchanged, whereas the second (Chapter 16) has been drastically revised in the direction of presenting the underlying models and their implication for the proper error term, or denominator for  $F$ . Assumptions are made more explicit.

Some will be critical of Chapter 18 because it does not contain all the several so-called nonparametric techniques. My choice from among them was based in part on computational facility

and lack of dependence on special tables. Furthermore, I have been unable to find clear-cut information concerning the relative efficiency of the many proposed techniques. Until such time as the mathematical statistician succeeds in evaluating their relative merits it seems unwise to confront the student with an aggregation of distribution-free methods.

As in the first edition, I have aimed for conciseness, with stress on assumptions and interpretations instead of on routine computational procedures. Some derivations have been included for the dual purpose of clarifying concepts and stimulating the mathematically inclined.

It is impossible to disentangle and acknowledge all the factors that have contributed to the content and writing of the first edition and this revision. Some will perhaps recognize the influence of two of my teachers, Professors Truman L. Kelley and Harold Hotelling. My greatest personal indebtedness is to Olga W. McNemar, who has done much to clarify the exposition and rid the volume of errors.

I am indebted to Professor Ronald A. Fisher and Dr. Frank Yates, also to Messrs. Oliver and Boyd Limited, Edinburgh, for permission to reprint Tables III, IV, V, and VII from their book "Statistical Tables for Biological, Agricultural and Medical Research."

QUINN McNEMAR

*Palo Alto*  
*July, 1954*



## Contents

1 · Introduction	1
2 · Tabular and graphic methods	5
3 · Describing frequency distributions	13
4 · Distribution curves	32
5 · Probability and hypothesis testing	41
6 · Inference: Continuous variables	72
7 · Small sample or $t$ technique	104
8 · Correlation: Introduction and computation	115
9 · Correlation: Interpretations and assumptions	122
10 · Factors which affect the correlation coefficient	144
11 · Multiple correlation	169
12 · Other correlation methods	191
13 · Frequency comparison: Chi square	212
14 · Comparison of variabilities	243
15 · Analysis of variance: Simple	249
16 · Analysis of variance: Complex	281
17 · Analysis of variance: Covariance method	343
18 · Distribution-free methods	357
19 · Remarks on error reduction	361
Exercises	367
Appendix (Tables)	381
Index	401



## Introduction

Statistical methods are concerned with the reducing of either large or small masses of data to a few convenient descriptive terms and with the drawing of inferences therefrom. The data are collected by any of several methods of research with the aid of measuring devices appropriate to a given area of investigation. The research methods are variously named and classified. Thus in psychology we have methods which are labeled experimental, clinical, observational, etc. The devices for measuring or securing responses vary from those which involve delicate apparatus through paper-and-pencil schemes to controlled observations and interviews. Statistical techniques are not to be considered as coordinate either with research methods or with devices for obtaining and recording responses, but rather as tools for analyzing data collected by whatever means.

The reduction of a batch of data to a few descriptive measures is the part of statistical analysis which should lead one to a better over-all comprehension of the data. All readers will be more or less familiar with the concept of *average*. An average is a measure which describes what is typical of a group with respect to some trait, characteristic, or variable. If one is comparing two or more groups, the determination of an average for each group permits a better appraisal of possible group differences than would be obtained by casual examination of the data. There are various statistical measures, or types of averages, which have proven useful as descriptive terms for a variety of data. One aim of this book is to present and discuss the descriptive statistical measures most frequently needed in psychological research. Proper usage and interpretation of these terms and evaluation of their use by others are not possible without knowledge of their meaning and



their limiting assumptions. Incidentally, the user of statistical measures must give some thought to computational procedures.

As we proceed it will be necessary not only to define descriptive measures but also to distinguish between the usage of a given measure as being descriptive of a *sample* as opposed to a *population*. Since sample descriptive statistics are known (i.e., computable) whereas the corresponding population values are unknowns (but estimable), we will in this book define and discuss the descriptive measures in terms of samples and subsequently consider the problem of drawing inferences about, or estimating, population values. Sample values are frequently referred to as *statistics* and population values are called *parameters*.

That part of statistical analysis which has to do with the drawing of inferences is imposed upon us because of certain inadequacies of research data. For instance, an investigator who wishes to know the average height of adult women in the United States will never have facilities for measuring every woman. Accordingly, he is compelled to measure a sample of women; then on the basis of information yielded by the sample he can make an inference concerning the average height of the population of women. Another investigator, wishing to evaluate the relative merits of 2 learning methods, tries out the methods with 2 small groups of students, and from the results an inference is made concerning what might be expected if he had facilities for working with very large groups. An opinion poller may seek information about the reactions of Republicans and Democrats to some world event. By questioning a sample of each group he can secure sufficient data for drawing an inference regarding a possible difference between the population of Republicans and the population of Democrats.

The problem of statistical inference is usually that of determining whether statistical significance can be attached to results after due allowance is made for known sources of error. There are many and varied situations for which we need tests of significance, and accordingly several tests are available. Intelligent and critical inferences cannot be made by those who do not understand the purposes, assumptions, and applicability of the various techniques for judging significance.

It is in connection with the problem of drawing inferences that a knowledge of statistical methods is most helpful. A research

should be planned in such a way that the resulting data are amenable to treatment by the available statistical techniques. With sufficient information concerning these techniques of analysis, one should be able to lay out in advance of data collecting the main types of statistical analysis to be used. If a proposed experimental setup precludes the possibility of adequate analysis, it may be found that a slight alteration in the plan will remedy the situation. All too frequently the statistician is called in to help with data which have not been collected in such a manner as to permit efficient analysis. Only by knowing the available methods of analysis can one plan a research with assurance that the results can be handled statistically.

Another reason for keeping in mind statistical considerations while planning a research is the fact that some experimental designs are preferable because they permit, with small additional cost, or even at a saving, better control of error than other plans. Indeed, certain designs lead to a marked reduction in known sources of error.

A third reason for planning with foresight regarding the statistical analysis is that a set of data can sometimes be made to serve for checking several different hypotheses.

The student should be warned that he cannot expect miracles to be wrought by the use of statistical tools. Although statistical methods have an important place in present-day psychological research, it does not follow that they can be utilized to salvage data that result from a haphazardly planned and sloppily executed investigation. No amount of statistical juggling can transfigure bad data into acceptable form. It is doubtful whether the student who comes to the statistician with a batch of data and the question, "Can I compute a correlation coefficient . . . ?" will make a scientific contribution, but such a student deserves sympathy, especially if his major advisor has suggested that he need not worry about statistics until he has collected data.

The purpose of the present book is to acquaint the student with the statistical techniques commonly used, to suggest economical computational procedures, and to state the assumptions and limitations of the various techniques. Whenever the understanding of a particular technique can be clarified by a simple derivation, such a derivation will be given. Unfortunately, many of the derivations are too complicated mathematically to permit con-

sideration in an elementary or intermediate treatment. The qualified and interested student will find some of these derivations in more advanced textbooks and others in original sources.

Statistical methods belong in the realm of applied mathematics, and consequently extensive scholarship in mathematics is required of those who choose to specialize in statistics. One can, however, secure a practical working knowledge of statistical techniques without first becoming a mathematician, provided his deficiency in mathematics is not accompanied by an emotional reaction to symbols.

Within the realm of psychological research there is wide variation in the need for statistical procedures. One can find current research reports which involve no use of statistics, some which involve very simple statistical treatment, still others which lean heavily on the tools of statistics, and a few which are highly statistical. One need not shift from one area of investigation to another to find this variation, but it is true that certain areas of research in psychology have less dependency than others on statistical procedures. The area of psychology which seems most dependent upon statistics is psychological measurement. This dependency is due mainly to the very nature of psychological measurement, the theory of which is largely statistical.

The presence or absence of statistical analysis *per se* is not a safe criterion for judging the worth of a study—some studies would have been improved by the utilization of statistics, whereas others would be better if they had been so designed as to depend less upon statistical analysis. Except for the requirement that the statistical analysis be adequate, there are no general rules as to how statistical a research should be. Of 2 experimental plans, either of which would provide appropriate data for checking a given hypothesis or sets of hypotheses, that plan which calls for simple statistical analysis is certainly preferable to the one which requires elaborate analysis. Experimental control of errors is far better than statistical adjustments.



## CHAPTER 2

### Tabular and Graphic Methods

When we are faced with a mass of data, the first manipulative step is tabulation or classification. If we are dealing with the number of children per family, the tabulation is equivalent to counting the number of one-child families, two-child families, etc.; or if we have information on 1000 persons regarding their national origin, we can tabulate, or count, the number of those of German, French, Italian, etc., origin; or these same individuals can be classified as to eye color. If we have their heights, we can also classify (or tabulate) them as being 58, 59, 60, etc., inches in height, and if the shortest person is 58 and the tallest is 78 inches, we would tabulate our 1000 into 21 different inch groups. If we also know the weights of these individuals, we can classify again, this time as 100, 101, up to (say) 229 pounds, and thereby have 130 groups. In all these situations we can classify with respect to the given characteristics, but the resulting tabulations will show marked differences as we pass from trait to trait. For instance, we may have only six national groups, and it will make little difference whether Germans or Russians are first on the tabulation sheet. Such a characteristic as nationality or eye color is said to be *unordered* (and somewhat *discrete*). The number of children per family is discrete but can be ordered, from least to greatest number. Now such a trait as height can also be ordered, but it is said to be *continuous* (nondiscrete) because it is possible to have an infinite number of in-between values very closely spaced. Such a series is sometimes called *graduated*. It will of course be obvious that a discrete series does not permit of in-between values, e.g., no family can have  $2\frac{1}{4}$  children.

For most purposes it is adequate if we tabulate, or classify, individuals into certain large groups. For example, instead of classifying our 1000 persons into pound groups (130 such groups)

it is usually sufficient to classify them into broader groups, say 100-109, 110-119, etc., thereby obtaining 13 large groups. As a matter of fact, the use of fewer groups has a distinct advantage in that the labor of tabulating and computing descriptive terms is greatly lessened. The factors influencing the choice of the grouping interval are two: first, its size should be such as to permit at least 10 or 12, but not more than 20, classes or groups; and second, it should promote tabulating convenience. Suggestions for choosing tabulating intervals are: (1) determine the *range* of measures or scores, i.e., the difference between the lowest and highest; (2) by inspection determine whether the range can be divided into 12 to 20 equal intervals of some convenient size, say 5 or 10; and (3) let the lower number of each interval be a multiple of the size of the interval. It is customary to arrange the tabulation sheet with the highest or largest values of the variable at the top and to use either dots or tally marks when tabulating. The tallies per interval can be counted and recorded to the right of the tally marks. This column is usually labelled  $f$ , and the sum of the  $f$ 's will be  $N$ , or the total number of individuals in all the grouping intervals. Tabulation results in a *frequency table* or *frequency distribution*, such as that shown in the first two columns of Table 1.

Table 1. FREQUENCY DISTRIBUTION OF IQ'S FOR 161 FIVE-YEAR-OLD BOYS

Interval	$f$	Smoothed $f$	Cumulative $f$
160-169	1	.3	161
150-159		1.3	160
140-149	3	4.0	160
130-139	9	13.7	157
120-129	29	25.7	148
110-119	39	34.3	119
100-109	35	35.3	80
90- 99	32	25.0	45
80- 89	8	14.0	13
70- 79	2	3.7	5
60- 69	1	1.3	3
50- 59	1	1.0	2
40- 49	1	.7	1

It should be noted that the expressed interval limits in a frequency table are not necessarily the actual limits. Thus, if weight has been taken to the nearest pound, the actual limits of the inter-

val 130-139 would be 129.5 and 139.5; but if the ages of individuals have been taken as at the last birthday, the interval 20-24 would have actual limits of 20 and 24.999+. Obviously for purposes of tabulation we need not use the implied actual limits, and for computational purposes we usually need either the lower limit or the midpoint of certain intervals, so there is nothing to be gained by meticulously labeling the intervals with actual limits.

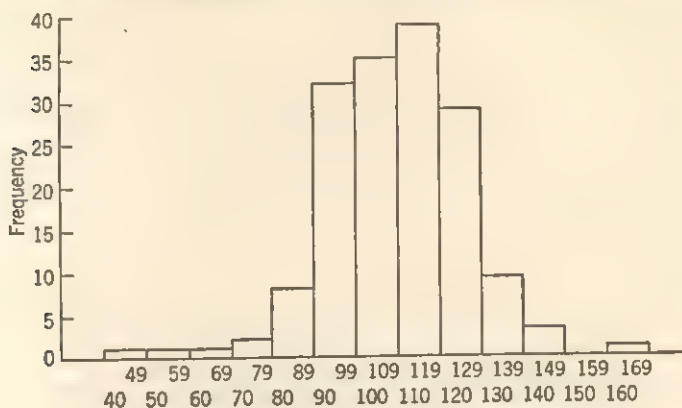


Fig. 1. Histogram for data of Table 1.

### GRAPHIC PRESENTATION

If one scrutinizes the tally marks or the frequency table, he can obtain some notion as to how the individual values are distributed. A number of pictorial schemes have been suggested as aids in the study of frequency distributions. It is possible to lay off the various values (or intervals) of the variable on the horizontal or  $x$  axis, and to let the vertical or  $y$  axis represent the frequency per value or interval. The frequencies of the several intervals can be represented by drawing a horizontal line across each interval at the height corresponding to the number of cases in that interval, and then connecting these horizontals with verticals erected at the interval limits. This yields a *histogram* (Fig. 1). Using the same arrangement of the vertical and horizontal scales, one can merely indicate the frequency with a dot or cross placed directly above the midpoint of the interval, and then connect the adjacent points with straight lines. This results in a *frequency*

*polygon* (Fig. 2). Such a polygon or the corresponding histogram will usually show irregularities; on the assumption that these are due to the operation of chance, one can draw a smooth curve, cutting as near the points as possible, and this curve can be thought of as giving a better picture than the original polygon. A curve which is obtained by freehand drawing or by graphic smoothing schemes or by repeated smoothing of the frequencies by a method of moving averages is known as a *frequency curve*. One method of

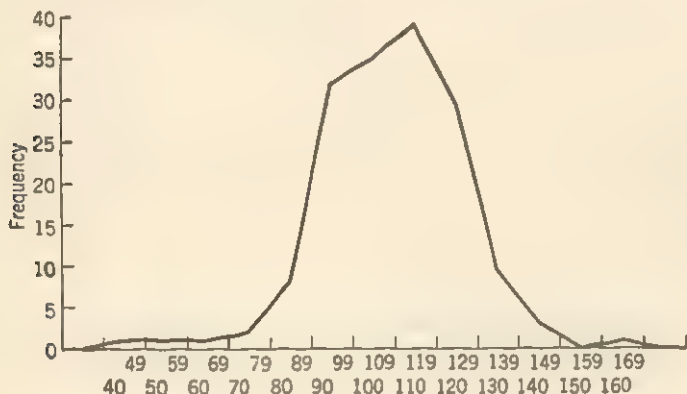


Fig. 2. Frequency polygon for data of Table 1.

moving averages is illustrated in Table 1, in which an average is taken over 3 intervals. The smoothed value for an interval is obtained by summing the frequencies in that interval and the 2 adjacent intervals and dividing by 3. Thus the smoothed value for the interval 80-89 is equal to the sum of the frequencies 2, 8, and 32, divided by 3. For the 90 interval, 8, 32, and 35 are summed and divided by 3. The student should plot both the original and smoothed frequencies so as to compare the 2 graphs.

Although it is relatively easy to depict a frequency distribution by a histogram or by a frequency polygon or by a smoothed frequency curve, it is necessary that we note a shift in interpretation as we pass from the histogram to the polygon to the curve. In drawing the histogram we are in effect drawing a series of vertical bars with a common boundary for any 2 that are adjacent to each other. Since the height of each bar represents a frequency, we may, by arbitrarily assigning unity as the width of each bar, say that the



area of a bar also represents a frequency. Then the sum of the areas of the several bars will be the total number of cases, or  $N$ .

If we think of the polygon in Fig. 2 as being superimposed on the histogram of Fig. 1 and imagine that the common boundaries of the vertical bars have been erased, we will have a picture like that in Fig. 3, in which the remaining parts of the bars have the appearance of an up and then down irregular staircase. A little thought

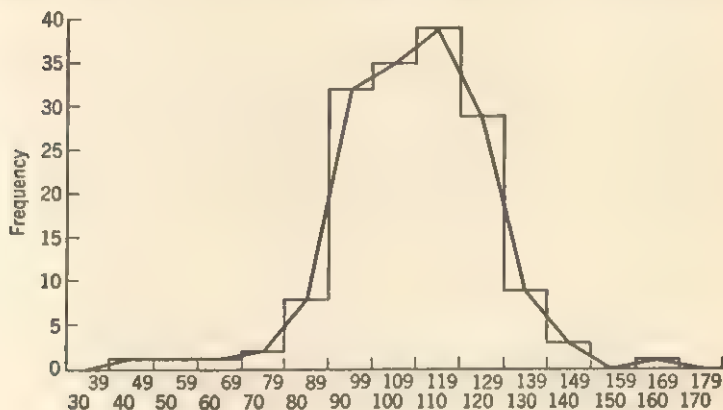


Fig. 3. Frequency polygon superimposed on histogram.

should convince the reader that the total area under this staircase is  $N$ , or precisely the same as the sum of the areas of all the bars.

Next consider the polygon. Note that as we pass from interval to interval the polygon in conjunction with the staircase histogram forms a series of pairs of equal-area triangles. One of each pair is an area included under the polygon but not under the histogram, while the other is an area included under the histogram but not under the polygon. The net effect of this balancing of areas, in and out, is that the total areas under the polygon and histogram are equal; each total area represents  $N$ .

Now it should not stretch one's imagination too much to regard the total area under a smoothed polygon or under a frequency curve as being equal to  $N$ . With this notion that area, not height, represents frequency we can readily speak of the area under the curve between ordinates erected at any 2 score values on the base line ( $x$  axis) as the number of cases between the 2 score points.

And of course the area under any part of the curve could be expressed as a proportion or a percentage of the total area.

This concept of area as frequency will have considerable value for us as a basis for interpreting certain statistical measures, and the concept will be indispensable to our understanding of certain "ideal," or mathematical, frequency curves, as yet undefined.

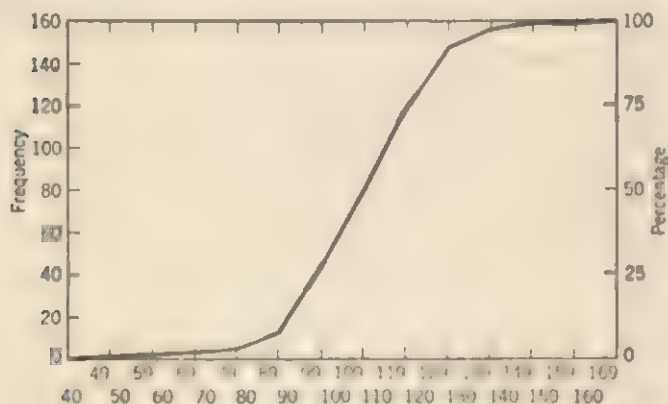


Fig. 4. Ogive for data of Table 1.

Another type of graph can be obtained by the use of *cumulative frequencies*. In Table 1 will be found a column headed "Cumulative *f*." These values are obtained by successive adding of the frequencies, beginning with the lowest interval. Adding 1 and 1 gives 2, adding to this the next frequency gives 3, to which in turn is added the next, giving 5, and so on until we have 160 plus 1 for the last cumulative value, which is the total number of cases. Obviously, from the cumulative table one can tell how many individuals fall below a given point. If one plots the cumulative values and connects the plotted points, an *ogive* curve results (Fig. 4). Note that in plotting the cumulative frequencies, one does not use the midpoint of the interval, but rather the upper boundary. Why?

The use of frequency polygons in the comparison of 2 groups is quite simple and often very enlightening. All that is necessary is to plot the data for both groups on the same sheet and with reference to the same axes. If the number of cases in the 2 groups

differs markedly, a better comparison can be obtained by converting the frequencies for each group to percentages of the total number in each group. Polygons based on percentage frequencies will not portray differences which are merely a reflection of differing  $N$ 's and therefore are more comparable. A glance at 2 such frequency polygons will reveal whether the 2 groups show marked differences in the trait in question or to what extent the 2 distributions overlap. More refined methods for comparing groups will be discussed later.

When one wishes to picture a discrete series it is customary to use either horizontal or vertical bars, separated from each other, to represent the several frequencies. As in the case of frequency polygons and histograms, there are no hard and fast rules regarding the heights (or lengths) of the bars relative to the horizontal (or vertical) base. The student should attempt to avoid extreme lack of proportion. Often in newspapers and magazines one finds that frequencies have been represented as areas or solids. A circular diagram, or pie chart, in which the sizes of the separate sectors represent the percentage falling into given groups or classes is sometimes used to picture relative frequencies. There is some evidence, and a general consensus of opinion, that some type of linear graph is less likely to be misinterpreted than one depending upon areas or solids.

Another type of graphical representation is used to picture the relationship between 2 variables, e.g., growth in stature and age, or price change with year. To make such a line graph, one can lay off time or age or trials, on the horizontal axis, choose a convenient scale on the  $y$  axis for the other variable and then plot the observational values. The line graph should be arranged so that the graph is read from left to right and from the bottom to the top, and the scales on the 2 axes should allow the inclusion of all observed values of the 2 variables and at the same time permit of a well-balanced or well-proportioned picture. A line graph can be made misleading by the choice of the scales on the 2 axes. For instance, if one is plotting the practice curve for card sorting (number of cards sorted on  $y$  axis, trial number on  $x$  axis), it is possible to make a tremendous difference in the appearance of the graph simply by altering the scale on the  $y$  axis. On 2 curves which represent the same relationship, one (Fig. 5a) would give the impression that the learning had progressed quite rapidly,

These are shown in the accompanying figures. The curves are plotted on a logarithmic scale, and the values are given in the accompanying table.

The curves are the result of a series of experiments, and the values are given in the accompanying table.

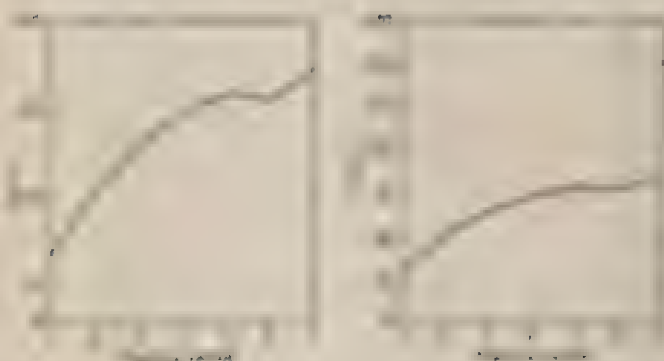


Fig. 1. Graph of temperature vs. time. Fig. 2. Graph of temperature vs. time.

These curves are the result of a series of experiments, and the values are given in the accompanying table. The curves are plotted on a logarithmic scale, and the values are given in the accompanying table.

The curves are the result of a series of experiments, and the values are given in the accompanying table.

These curves are the result of a series of experiments, and the values are given in the accompanying table.





distributed scores. Since the computation of the descriptive terms frequently involves a determination of the lower limit or midpoint of a class interval, the student should recall what has been said about actual and expressed class limits. Obviously, if one needs the midpoint of an interval, it is necessary only to add one-half the size of the interval to the actual lower limit, which must be determined by a consideration of the nature of the scores or measures which constitute the variable. Psychological measurements and test scores are usually treated as though rounded to the nearest value.

### MEASURES OF CENTRAL VALUE

**The mode.** A glance at a typical frequency distribution will indicate to us the most frequently occurring  $X$  value, or for grouped data the group of  $X$  values which has the greatest frequency. This maximal frequency roughly defines the *mode*. For non-grouped data the mode is the  $X$  value having the greatest frequency, whereas for grouped data the mode is taken as the midpoint of the interval which has the greatest frequency. For a smoothed frequency curve, the mode is the  $X$  value at which the curve reaches its maximum height. The mode is one indicator of central value, but as a descriptive statistic it has serious limitations. If one uses a different size interval, the mode may be decidedly different. Furthermore, it occasionally happens that 2 nonadjacent intervals have the same maximal frequency, thereby yielding 2 modal values. Such a distribution is said to be bimodal, but it should be noted that the bimodality may not be real but merely accidental, the resultant of the particular grouping interval which has been chosen. In dealing with certain discrete series, like size of family, the modal value is apt to be more typical than some other measure of central value and therefore should be used, even though as a measure it is subject to greater sampling fluctuations than either the mean or the median. (The question of sampling cannot be discussed at this time; the student is asked to take on faith statements regarding the efficiency of a given statistic.)

**The median.** As a measure of central value, the *median* is defined in 2 ways: (1) if the individual scores are arranged in order with respect to some trait, the median is the value of the midmost individual if  $N$  is odd, or lies midway between the 2 middle in-

dividuals when  $N$  is even; (2) when a distribution has been made, the median is defined as the point on the scale such that the frequency above or below the point is 50 per cent of the total frequency. For grouped data, the median may be determined by the following steps:

1. Find one-half of  $N$ .
2. Count the frequencies in a cumulative manner from the bottom up to that interval, say the  $s$ th, the frequency of which if included would give more than, if not included less than,  $N/2$  cases. Obviously the median will fall somewhere in this interval unless exactly half the values fall below the lower limit of an interval, in which case this lower limit is the median. Let  $F_c$  equal the total frequency up to the  $s$ th interval, and let  $F_s$  equal the frequency in the  $s$ th interval.
3.  $(N/2 - F_c)/F_s$  will be the proportional distance required in the  $s$ th interval to locate the median.
4. Letting  $i$  equal the size of the interval and  $LL_s$  the lower limit of the  $s$ th interval, the median will be given by

$$\text{Mdn} = LL_s + i \frac{N/2 - F_c}{F_s} \quad (1)$$

This involves the defensible assumption that the scores for the cases falling in the  $s$ th interval are distributed fairly evenly over the possible score values in the interval.

The calculation of the median is illustrated in Table 2, in which is given the distribution of scores made by 50 college men on the

Table 2. THE CALCULATION OF THE MEDIAN

Score	$f$	
310-319	1	
300-309	2	
290-299	4	$N/2 = 25$
280-289	1	$s$ th interval is 280-289
270-279	6	$F_c = 24$ $F_s = 12$
260-269	12	$i = 10$
250-259	11	$LL_s = 259.5$
240-249	8	
230-239	2	$\text{Mdn} = 259.5 + 10 \frac{25 - 24}{12} = 260.33$
220-229	0	
210-219	3	
	50	

Brown spool packer. The score is the number of spools packed in four 1-minute trials.

The chief merits of the median are its ease of computation, its independence of extremes (it can be computed even if a known number of extremes have not been measured), and the fact that it is not affected by the size of extremes. This last point will be clearer after a discussion of the mean.

**The mean.** This arithmetic average will already be familiar to most readers. The *mean* is defined simply as the sum of all the scores or measures divided by their number or

$$M = \frac{\Sigma X}{N} \quad (2)$$

where  $X$  represents any score, the symbol  $\Sigma$  means "the sum of," and  $N$  is the total number of cases. When  $N$  is small, this definition form can be used to compute the mean, but when  $N$  is large, say 50, 100, or more, such a method is not economical of time. Ordinarily, when  $N$  is large, one makes a frequency distribution from which it is possible to compute the mean and median and other statistical measures. Assuming that the midpoint of an interval is typical of all the individuals in the interval, one can obtain the mean by summing the products of the several interval midpoints by their respective frequencies and dividing this sum by  $N$ . The error introduced by the use of midpoints is nonsystematic, i.e., tends to be ironed out so far as the computed mean is concerned.

The computation of the mean can be shortened further by use of an arbitrary origin and deviations therefrom. The reasonableness of such a procedure can be readily grasped by considering the problem of determining the mean height of a group of men. We could measure each man's height from the floor or as so much in excess of a stationary bar 5 feet from the floor. The sum of the excesses divided by  $N$  will be the mean excess, and obviously we must add 5 feet to this to obtain the mean height of the group.

When we have a frequency distribution the arithmetic can be shortened still further by expressing the deviation from an arbitrary origin in terms of step intervals, that is, as the number of intervals that a given interval deviates from the arbitrary origin. The arbitrary origin is taken as the midpoint of any interval, and



it is assumed that the midpoint of each interval may be taken as representing the scores in that interval.

The procedure can be developed by simple algebra. Let  $AO$  be the arbitrary origin,  $i$  be the interval size, and  $d$  be the deviation in step intervals of the midpoint of any interval from  $AO$ . Then each score can be expressed as  $X = AO + id$  in which  $AO$  and  $i$  are constant and  $d$  varies. From the definition formula for the mean we have

$$M = \frac{\Sigma X}{N} = \frac{\Sigma(AO + id)}{N} = \frac{\Sigma(AO) + \Sigma id}{N}$$

Now  $\Sigma(AO)$  will equal  $N(AO)$  because summing a constant  $N$  times is the same as multiplying it by  $N$ . As an exercise, the student should demonstrate, by taking varying numbers each multiplied by a constant, that  $\Sigma id = i\Sigma d$ ; a constant can be brought out from under the summation sign. Hence we have

$$M = \frac{N(AO)}{N} + \frac{i\Sigma d}{N} = AO + i \frac{\Sigma d}{N}$$

Since we started by summing  $N$   $X$ 's and since each  $X$  is associated with a  $d$  value, we should be summing  $N$   $d$ 's. That is, the  $d$  value for a particular interval needs to be summed  $f$  times ( $f$  being the frequency for the interval), but the sum for a particular interval is simply  $f$  times its  $d$ . If we replace  $\Sigma d$  by  $\Sigma fd$  we explicitly indicate that each  $d$  is to be summed as often as it occurs. Accordingly, our computational formula for the mean is written as

$$M = AO + i \frac{\Sigma fd}{N} \quad (3)$$

In our algebraic derivation of formula (3) the only restriction placed on  $AO$  was that it be the midpoint of an interval; hence we are free to choose arbitrarily the midpoint of any interval as  $AO$ . In order to avoid negative  $d$ 's,  $AO$  is ordinarily taken as the midpoint of the lowest interval. Table 3 indicates the computation of the mean from grouped data by use of an arbitrary origin and deviations therefrom in terms of step intervals.

Table 3. CALCULATION OF THE MEAN

Score	$f$	$d$	$fd$	
310-319	1	10	10	
300-309	2	9	18	
290-299	4	8	32	
280-289	1	7	7	$\Sigma fd = 235$
270-279	6	6	36	
260-269	12	5	60	$i \frac{\Sigma fd}{N} = 47.00$
250-259	11	4	44	
240-249	8	3	24	
230-239	2	2	4	$M = 214.5 + 47.00 = 261.50$
220-229	0	1	0	
210-219	3	0	0	
	<hr/> 50		<hr/> 235	

If we had taken  $AO$  near the center of the distribution we would be following the so-called *guessed average* method, a method which has the advantage of smaller  $d$  values but has the disadvantage of both negative and positive  $d$ 's.

Parenthetically, it might be pointed out that the use of the arbitrary origin, step-interval scheme is analogous to using *coded* scores. If we regard  $d$  as a coded value we see from  $X = AO + id$  that  $d = (X - AO)/i$ , or that in general we have a coded score  $X_c = (X - K)/k$ , with  $K$  and  $k$  so chosen as to give coded values ranging from zero to between 10 and 20. Then the mean of the original scores is given by  $M = K + k$  times the mean of the coded scores.

The beginning student who is puzzled about which measure to use, the median or the mean, should remember that the purpose of measures of central value is description. When one is attempting to reduce a mass of scores or a distribution of measures to a few descriptive constants, the mean and median are both descriptive terms which more or less adequately depict the "average" or typical score, and the choice between the two is frequently determined on the basis of which is more typical. Thus, if 6 men run 100 yards in 9.6, 9.7, 9.8, 9.9, 10.0, and 11.0 seconds, the mean value of 10.5 is not as typical as the median value of 9.85. In general, the mean is not as typical as the median when there are extreme measures in one direction. However, when the scores are distributed in an approximately symmetrical fashion, the mean and median will be equal or nearly so, and either will be as typical

as the other. The mean in this case has 2 distinct advantages over the median: (1) It is usually a more stable measure in the sampling sense, i.e., if we regard our scores as based on a sample of  $N$  individuals and then take another sample, the means of the 2 samples will in general show closer agreement than the 2 medians. This point will be discussed in more detail in the chapter on sampling errors. (2) It can be handled arithmetically and algebraically. The student should prove that, if the mean of  $N_1$  cases is  $M_1$ , and of  $N_2$  cases is  $M_2$ , the mean of the 2 groups combined will be given by

$$M_c = \frac{N_1 M_1 + N_2 M_2}{N_1 + N_2}$$

The median cannot be handled in such a fashion. Furthermore, the mean is used in connection with more advanced topics in statistics, whereas the median is seldom mentioned. Thus, unless the distribution is markedly skewed, the mean should be used. The problem of describing skewness will receive consideration after measures of variation have been discussed.

As exercises, the student should show algebraically or to his own satisfaction by numerical examples that (1) if a constant is added to or subtracted from the scores of a group, the new mean will be  $M + C$  or  $M - C$ , where  $C$  is the given constant and  $M$  the mean of the original scores; (2) if all the scores are multiplied by a constant,  $C$ , the new mean will be  $C M$ , whereas dividing by a constant will lead to  $M/C$  as the new mean.

## MEASURES OF VARIATION

The description of the extent of scatter (or cluster) about the central value may be obtained by any one of several measures. These measures differ somewhat in interpretation and usefulness. One may doubt whether the *range* (highest to lowest score) is of sufficient value in psychological research to justify its use as a measure of variation. It is, obviously, determined by the location of just 2 individual measures or scores and consequently tells us nothing about the general clustering of the scores about a central value.

**Quartile deviation.** An easily computed description of dispersion is the *quartile deviation* ( $Q$ ), defined as  $(Q_3 - Q_1) / 2$ , in

which  $Q_3$  (or the third quartile) is the point above which one-fourth of the cases fall and  $Q_1$  (or the first quartile) is the point with three-fourths of the cases above.  $Q_2$  (or the median) has already been defined as the point above which one-half of the cases fall. The computation of the 2 quartiles  $Q_3$  and  $Q_1$  from grouped data is essentially the same as that of the median. For instance, in determining the third quartile we count up to the interval in which the point falls which divides the number of cases into 2 parts: three-fourths below and one-fourth above. The distance into this interval is found in exactly the same manner as in computing the median. Since the quartiles are not influenced by extremes, it is customary to use them along with the median. By definition, 50 per cent of the cases fall between the first and third quartiles, but in nonsymmetrical distributions it is not likely that the limits indicated by the median plus and minus  $Q$  will include 50 per cent. It would seem better to report both the first and third quartiles, instead of  $Q$ , as these values along with the median will enable one to picture whether or not the clustering above the median is different from that below the median.

**Percentiles.** Closely allied to the quartiles are the percentiles. The  $P$ th percentile is defined as a point below which  $P$  per cent of the cases fall. Thus the median is the 50th, the third quartile the 75th, and the first quartile the 25th percentile. The 10th, 20th,  $\dots$  90th percentiles are sometimes called deciles. The computation of the percentiles from grouped data is accomplished in the manner indicated for computing the quartiles. The location of the zeroth and 100th percentiles is always perplexing. Since these 2 points are dependent upon the location of just 2 scores (i.e., are greatly influenced by chance), they are difficult to interpret. Common sense would suggest that the concept of these 2 percentiles be dropped.

Percentiles may readily be associated with the cumulative frequency distribution, and with the ogive curve if cumulative percentage frequencies (obtained by dividing the  $f$ 's by  $N$ ) are used along the ordinate when plotting the ogive. In fact, the ogive may be used as a graphic scheme for determining score values corresponding to given percentiles. For instance, if we wish to obtain the 25th percentile point, we find 25 on the ordinate scale, proceed horizontally to the ogive curve, then vertically to the  $x$  axis, and read off the score corresponding to the 25th percentile. Scrutiny

of Fig. 4, p. 10, will help the student understand the process. Could we also use the ogive as a basis for determining the percentile value of a given score?

The use of the difference between percentiles as an indication of dispersion should be obvious. In fact, the 10th 90th percentile range is a somewhat better (more stable from sample to sample) measure of dispersion than the quartile deviation. Percentiles, however, are chiefly of value in reporting the scores of individuals on psychological and educational tests. Ordinarily a raw score gives no inkling of what it means, whereas when it is said that an individual scores at or near the 85th percentile, the implication is that 15 per cent of his fellows score higher or better than he. Thus a percentile score carries with it some idea of the location of the individual with reference to the group. Furthermore, percentile scores for entirely different tests are comparable if derived from the same group or sample. The original raw scores might be different units, e.g., number of additions per minute and time to read a page of prose, and consequently not at all comparable.

**The average deviation.** Sometimes called the mean deviation or mean variation, the *average deviation* ( $AD$ ) is defined as the average of the deviations of the several scores from the mean. Thus, if  $x = X - M$ , then  $AD = \sum |x| / N$ , where  $|x|$  is the absolute value of  $x$ , i.e., the negative deviations are treated as though positive. Currently the average deviation is seldom used; the student, however, needs to know something about it if he reads the earlier research literature in psychology.

Contrasted with the quartile deviation, the average deviation gives weight to extremes, and for the usual bell-shaped distribution the limits  $M$  plus and minus  $AD$  will include about 57.5 per cent of the cases; the average deviation is larger than  $Q$  but not so large as the standard deviation, to which we now turn.

**The standard deviation.** A third measure of variation, the *standard deviation* ( $SD$  or  $\sigma$ ), is defined as

$$\sigma = \sqrt{\frac{\sum x^2}{N}} \quad (4)$$

where  $x = X - M$ . To compute the standard deviation directly from this formula would be very cumbersome and uneconomical, since  $x$  will usually involve decimals. A computational formula



involving deviations from an arbitrary origin ( $AO$ ) can be easily derived by algebra. Such a derivation is included here in order further to familiarize the student with the method of handling summation signs. The derivation will be carried through for  $\sigma^2$ , technically known as the *variance*; then at the end we can take the square root to obtain  $\sigma$ .

From formula (4) we have

$$\sigma^2 = \frac{\Sigma x^2}{N}$$

in which  $x = X - M$ .

As in deriving formula (3), we can set

$$X = AO + id$$

and since  $M = AO + i(\Sigma d/N)$ , we have, substituting in  $x = X - M$ ,

$$\begin{aligned} x &= AO + id - \left( AO + i \frac{\Sigma d}{N} \right) \\ &= id - ic \end{aligned}$$

where for convenience we let  $c$  stand for  $\Sigma d/N$ .

$$\begin{aligned} x^2 &= (id - ic)^2 = i^2(d - c)^2 \\ \Sigma x^2 &= i^2 \Sigma (d - c)^2 \\ &= i^2 (\Sigma d^2 - 2c \Sigma d + Nc^2) \end{aligned}$$

Dividing both sides by  $N$ , we have,

$$\begin{aligned} \sigma^2 &= \frac{\Sigma x^2}{N} = i^2 \left( \frac{\Sigma d^2}{N} - 2c \frac{\Sigma d}{N} + N \frac{c^2}{N} \right) \\ &= i^2 \left[ \frac{\Sigma d^2}{N} - 2 \left( \frac{\Sigma d}{N} \right)^2 + \left( \frac{\Sigma d}{N} \right)^2 \right] \\ &= \frac{i^2}{N^2} [N \Sigma d^2 - (\Sigma d)^2] \end{aligned}$$

hence

$$\sigma = \frac{i}{N} \sqrt{N \Sigma d^2 - (\Sigma d)^2}$$

But since this form does not make explicit the fact that each  $d$ , and each  $d^2$ , must be summed as often as it occurs, we will insert  $f$

for the frequency of occurrence. Thus our computational formula becomes

$$\sigma = \frac{i}{N} \sqrt{N \sum fd^2 - (\sum fd)^2} \quad (5)$$

where  $\sum fd$  = the algebraic sum of deviations (in step intervals) from an arbitrary origin, and  $\sum fd^2$  = the sum of the squares of the deviations (in step units). The arbitrary origin may be taken as the midpoint of the lowest interval or as a guessed average near the center of the distribution. The advantage of the latter procedure is that the  $d$ 's will be relatively small and consequently will not lead to the handling of large numbers, whereas the first procedure avoids the use of negative numbers and is more readily adaptable to machine computation.

The computation of  $\sigma$  for grouped scores is illustrated in Table 4, which is identical to Table 3 except that we now have an  $fd^2$

Table 4. COMPUTATION OF SD BY USE OF AN ARBITRARY ORIGIN

Score	<i>f</i>	<i>d</i>	<i>fd</i>	<i>fd</i> <sup>2</sup>	
310-319	1	10	10	100	
300-309	2	9	18	162	
290-299	4	8	32	256	
280-289	1	7	7	49	By formula (5):
270-279	6	6	36	216	
260-269	12	5	60	300	$\sigma = \frac{10}{50} \sqrt{50(1339) - (235)^2}$
250-259	11	4	44	176	
240-249	8	3	24	72	$= 21.66$
230-239	2	2	4	8	
220-229	0	1	0	0	
210-219	3	0	0	0	
	—		—	—	
	50		235	1339	

column. It is easily seen that the  $fd^2$  values can be obtained by multiplying the  $fd$  values by the corresponding  $d$ 's. If we regard  $d$  as a coded score ( $= X_c$ ) with  $i$  as the constant  $k$ , we see that (5) is appropriate for computing  $\sigma$  by way of coded scores as defined on p. 18.

The  $fd$  and  $fd^2$  columns need not appear on the work sheet when we are computing the mean and standard deviation by a Monroe or Marchant or Friden type calculating machine. The 2 required sums can be obtained by punching in the lowest  $d$  in the right-hand

part of the keyboard and the corresponding  $d^2$  just left of the center of the keyboard, multiplying both simultaneously by the given frequency, and then, without clearing the lower dial, punching in the next larger  $d$  and its square, and so on. The successive products so obtained will be accumulated by the machine so that  $\Sigma fd$  is read directly from the right-hand side of the lower dial, and  $\Sigma fd^2$  is read from near the center of the same dial. If either an 8- or 10-bank machine is used, the  $d$ 's of 9 and less are punched in the right-hand column of the keyboard, and higher values will of course require the first 2 columns. The squares of the  $d$ 's will ordinarily be less than 400, rarely greater than 961, so that their values can be punched in columns 6, 7, and 8. The student should note that the squares of 1, 2, and 3 are to be punched in column 6, the squares of 4 to 9 in columns 6 and 7, and the squares of 10 to 31 in columns 6, 7, and 8. The sum of the squares will appear in the lower dial from window 6 to the left. With a little practice the 2 required sums for a distribution of 15 intervals and 200 cases can be obtained in less than a minute. It should not be necessary to say that the computation should be done twice as a check.

For use with a calculator, formula (5) has an advantage over formulas which involve 2 divisions under the radical. Thus we place the sum of the squares in the right-hand side of the keyboard, multiply by  $N$ , and leaving the product in the lower dial, punch the sum of the  $d$ 's in the keyboard and subtract it  $\Sigma fd$  times, and then from the dial copy the value of  $N\Sigma fd^2 - (\Sigma fd)^2$ .

Briefly summarizing, it will be noted that (1) with a machine  $\Sigma fd$  and  $\Sigma fd^2$  taken from an arbitrary origin at the bottom of the distribution are no more difficult to compute than when taken from a guessed average, (2) all sums are positive, and (3) the 2 sums necessary for determining both the mean and standard deviation can be obtained in the same operation. It is helpful to write the  $d$  column in red on the work sheet, thereby throwing it into contrast with the  $f$  column.

When  $N$  is small and the scores are not too large,  $\sigma$  can be computed economically by way of the original (raw) scores. The definition formula, (4), calls for  $\Sigma x^2$ . Note that since each  $x = X - M$ , we have

$$\Sigma x^2 = \Sigma (X - M)^2 = \Sigma X^2 - 2M\Sigma X + \Sigma M^2$$

Replacing the last  $\Sigma$  by  $N$  (we are summing  $M^2$   $N$  times) and replacing  $M$  by  $\Sigma X/N$ , we have

$$\begin{aligned}\Sigma x^2 &= \Sigma X^2 - 2 \frac{\Sigma X}{N} \Sigma X + N \left( \frac{\Sigma X}{N} \right)^2 \\ &= \frac{N \Sigma X^2 - 2(\Sigma X)^2 + (\Sigma X)^2}{N} \\ \Sigma x^2 &= \frac{1}{N} [N \Sigma X^2 - (\Sigma X)^2] \quad (6a)\end{aligned}$$

Substituting in formula (4) leads to an  $N^2$  in the denominator, which can be brought out as  $1/N$ , hence we have

$$\sigma = \frac{1}{N} \sqrt{N \Sigma X^2 - (\Sigma X)^2} \quad (6b)$$

All the scores are simply squared and then summed to get  $\Sigma X^2$ , and  $\Sigma X$  has the same meaning as in formula (2).

Although a mean computed by formula (2) from grouped data will not err systematically from the value obtained by formula (1), the use of formula (5) for calculating  $\sigma$  tends to give a value which is too large when compared with the nonapproximate value yielded either by (4) or by (6b). The reason for this is easily explained at the blackboard—we give here a hint. In general for an interval below the mean there will be more scores above than below the midpoint of the interval, while for an interval above the mean there will be more scores below than above the midpoint. Thus in taking the several midpoints as representing the scores within the several intervals we are in effect using values which deviate too far from the mean.

We may correct for the systematic error involved in using formula (5) by substituting in

$$\sigma_c = \sqrt{\sigma^2 - \frac{i^2}{12}} \quad (7)$$

The  $i^2/12$  is known as Sheppard's correction for grouping. The uncorrected and corrected values differ but little when 12 or 15

intervals have been used, and as the number of intervals is increased, the difference becomes smaller and smaller. If less than 10 intervals have been used, the error may be appreciable and the correction should be applied. These considerations form the basis for the suggested rule that at least 10 or 12, and not more than 20, intervals be used.

Regarding the interpretation of the standard deviation, it can be said that, when we have the usual symmetrical bell-shaped distribution, about 68 per cent of the cases will fall between the limits plus and minus  $1\sigma$  from the mean, about 95 per cent between plus and minus  $2\sigma$ , and nearly all the cases (99.73 per cent) between plus and minus  $3\sigma$ . The standard deviation, even more than the average deviation, gives weight to extremes and therefore may not be as good as the quartiles for describing the dispersion. The standard deviation has decided advantages over other measures of dispersion: (1) Typically, it is more stable from the sampling viewpoint. (2) It can be handled algebraically, i.e., if we have 2 groups of  $N_1$  and  $N_2$  cases, with  $M_1$  and  $M_2$ , and  $\sigma_1$  and  $\sigma_2$ , as the respective means and standard deviations, we can obtain the standard deviation for the 2 groups combined by

$$\sigma_c = \sqrt{\frac{N_1(M_1^2 + \sigma_1^2) + N_2(M_2^2 + \sigma_2^2)}{N_1 + N_2}} - M_c^2 \quad (8)$$

where the subscript  $c$  refers to the combined group. The mean for the combined group can be obtained by a formula given on p. 19. Formula (8) can be extended for determining the standard deviation for 3 or more groups combined. The student can make this extension as an exercise. (3) The standard deviation is a mathematical term which has considerable importance in more advanced statistical work. It is usually involved in the determination of sampling errors and is the measure of variation used in the analysis of variation and in connection with correlational analysis. Therefore, unless there are definite reasons for *not* using it, the standard deviation, instead of the average deviation or  $Q$ , should be used as a description of the amount of dispersion.

As an exercise, show that, if a constant is added to or subtracted from each of a set of scores, the standard deviation does not change, and that multiplying or dividing each by a constant will lead to



$C\sigma$  or  $\sigma/C$ , respectively, as the new standard deviation, where  $\sigma$  stands for the sigma of the original scores and  $C$  is the constant.

### MEASURES OF SKEWNESS AND KURTOSIS

If a distribution is not of the symmetrical bell-shaped type, it is not sufficient for descriptive purposes to report only the mean and standard deviation. We also need a measure of the lack of symmetry, i.e., of *skewness*, and frequently it is desirable to describe the distribution still further by giving a measure which indicates whether the distribution is relatively peaked or flat-topped, i.e., a measure of *kurtosis*.

Skewness can be described roughly by a number of measures, such as the difference between the mean and median divided by the standard deviation, or in terms of quartiles or percentiles. If an adequate and stable description of skewness is desired and if a measure of kurtosis is also needed, a method based on moments is to be preferred.

The first 4 *moments* about the mean are defined as follows:

$$\left. \begin{aligned} u_1 &= \frac{\Sigma x}{N} = 0 \\ u_2 &= \frac{\Sigma x^2}{N} = \sigma^2 \\ u_3 &= \frac{\Sigma x^3}{N} \\ u_4 &= \frac{\Sigma x^4}{N} \end{aligned} \right\} \quad (9)$$

where  $x$  represents the deviation of each score from the mean of all the scores. For purposes of computation, the moments about an arbitrary origin can be determined, and then from these values we can obtain the moments about the mean. This procedure has already been employed in computing the standard deviation; i.e., we took deviations from an arbitrary origin. [The definition of the standard deviation, formula (4), was in terms of deviations from the mean.] If we use  $v$  to represent moments about an arbitrary origin, the first 4 moments about  $AO$  can be defined as

follows, where  $d$  is the score deviation from  $AO$  in step units:

$$\left. \begin{aligned} v_1 &= \frac{\sum fd}{N} \\ v_2 &= \frac{\sum fd^2}{N} \\ v_3 &= \frac{\sum fd^3}{N} \\ v_4 &= \frac{\sum fd^4}{N} \end{aligned} \right\} \quad (10)$$

When the  $v$ 's have been calculated, the  $u$ 's can be readily determined from the following relationships:

$$\left. \begin{aligned} u_1 &= 0 \\ u_2 &= i^2(v_2 - v_1^2) = \sigma^2 \\ u_3 &= i^3(v_3 - 3v_2v_1 + 2v_1^3) \\ u_4 &= i^4(v_4 - 4v_3v_1 + 6v_2v_1^2 - 3v_1^4) \end{aligned} \right\} \quad (11)$$

The student should note the similarity of the formula in (11) for the second moment to that given for the standard deviation [formula (5)].

A *measure of skewness* defined in terms of moments is

$$g_1 = \sqrt{\beta_1} = \frac{u_3}{u_2\sqrt{u_2}} \quad (12)$$

For symmetrical distributions the value of  $g_1$  will be zero; hence the departure of  $g_1$  from zero can be taken as a measure of skewness. The deviation of  $g_1$  from zero, however, must be considered in light of the operation of chance or in terms of sampling errors (to be discussed later). The skewness is said to be positive when  $g_1$  is positive and negative when  $g_1$  is negative.

The *degree of kurtosis* can be described by

$$g_2 = (\beta_2 - 3) = \frac{u_4}{u_2^2} - 3 \quad (13)$$

When  $g_2$  is less than zero, the distribution tends to be flat-topped, whereas for  $g_2$  greater than zero it is relatively steep or peaked. When both  $g_1$  and  $g_2$  are zero or near zero, the distribution is of the usual symmetrical bell-shaped type, which is referred to as the "normal" frequency distribution.

Formulas (12) and (13) also define  $\beta_1$  and  $\beta_2$ , which have been and are still used as measures of skewness and kurtosis. Recently, the  $g$  measures have come into usage because of certain advantages which need not be discussed here.

It will be noted that the measure of skewness involves taking the third moment relative to  $\sigma^3$  (since  $u_2 = \sigma^2$ ), and that the measure of kurtosis depends upon the fourth moment relative to  $\sigma^4$ . For a given distribution, all the values of  $u_2$ ,  $u_3$ , and  $u_4$  are in terms of the same measurement unit, say inches or pounds or IQ's or minutes; hence the ratios in formulas (12) and (13) are pure numbers, i.e., are not inches or pounds or IQ's or minutes. If we have the distribution of the weights and of the heights for 1000 individuals, the measure of skewness for the height distribution may be compared directly with that for the weight distribution. This is true by virtue of the fact that for each we are expressing the third moment relative to the amount of variability, both in inches for one distribution, both in pounds for the other. Likewise, it can be reasoned that the measures of kurtosis for different distributions are comparable, although the distributions involve different measurement units.

In order to help the reader visualize the meaning of different values for  $g_1$  as associated with different degrees of asymmetry, Fig. 6 has been prepared.

When we have determined the mean and the second, third, and fourth moments, and from the moments have derived expressions which tell us the degree of dispersion, skewness, and kurtosis, we have a description which is adequate for most distributions. These measures can be used to determine the type of mathematical equation which will fit an observed frequency polygon; i.e., we can write the equation of a frequency curve which fits the observed frequency distribution. A distribution frequently found in psychological research is of the "normal" type, which is sufficiently described by the mean and standard deviation. Ordinarily it is not necessary to compute  $g_1$  unless the distribution "appears" to be skewed or to compute  $g_2$  unless the distribution seems peaked

or flat. The nature of the research, the type of variable being studied, and also the size of the sample are factors which need to be considered in making a decision as to the necessity for computing measures of skewness and kurtosis. It is seldom advisable to compute these measures when  $N$  is less than 100.

The student should be apprised of the fact that the rather frequent occurrence of symmetrical distributions for psychological

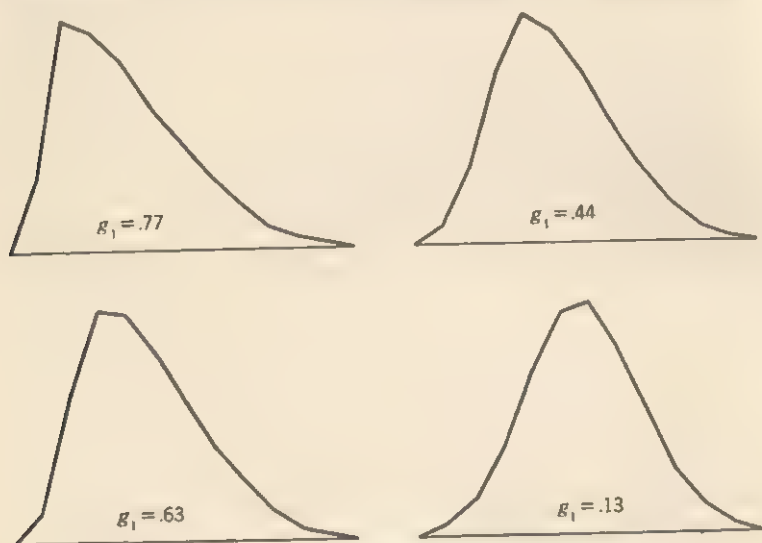


Fig. 6. Polygons with different degrees of skewness.

variables may result from an artifact, and also that the occurrence of a skewed distribution may likewise be artifactual. This is true because very few of the instruments used in psychological "measurement" involve equal unit scales—the measuring units are frequently arbitrary or even accidental. Many of the variables are measured simply in terms of the number of items checked or the number of items correct. The shape of the resulting distributions is largely determined by the percentage checking the items or by the difficulty of the items. If the items are of medium difficulty for a group, it can be expected that the scale will yield a symmetrical distribution when applied to the group; if the items are easy, the scores will pile up toward the top (give negative skewness); if difficult, a piling up toward the bottom will occur.

In the absence of equal scale units for the measuring devices one cannot really say whether the distribution of, for example, arithmetic ability for a given group is symmetrical or skewed—all that can be said is that in terms of the units used the distribution has a particular shape.

From the foregoing it would seem that, since skewness (and kurtosis too) is partly a function of the accidental nature of the measuring units, the descriptive measures of shape would have little value in psychology. The fact remains, however, that sometimes it is desirable to specify the skewness and kurtosis of a distribution of scores merely as a part of the description of what happens when a scale of measurement, however arbitrary the units, is applied to a given group. Furthermore, it is to the student's advantage to know something of measures of skewness and kurtosis because we shall later have occasion to refer to them, and because he is apt to encounter them in more mathematical treatments of statistics.



## CHAPTER 4

### Distribution Curves

By successive smoothing of a polygon (or distribution), one can iron out irregularities until the polygon becomes a "smooth" or regular and uniform curve. We can think of this curve as being similar or nearly identical to what we would obtain were we to increase indefinitely the size of our sample and at the same time use smaller and smaller grouping intervals. That is, the limit of a polygon, as we allow  $N$  to approach infinity and the interval size to approach zero, is conceived to be a curve which is smooth and regular. Now such a uniform curve can usually be described in terms of a mathematical equation. The student may recall that the general equation for a straight line is  $y = ax + b$ , and that  $y = 2x + 3$  is the equation for a particular line, that  $x^2 + y^2 = a^2$  is the equation for a circle of radius  $a$  with the origin or intersection of the abscissa and ordinate at the center, also that  $y = a + bx + cx^2$  is the general equation for a parabola. It is not until we give specific numerical values to the constants that we have equations for particular curves.

Frequency curves can be thought of as representing the relationship between two variables:  $y$ , or the height of the curve, and  $x$ , the variate or variable under consideration. Frequency polygons or distributions, even when smoothed, may be of various shapes: symmetrical or skewed, flat-topped or steep, humped near the center or at one end, bimodal or unimodal, J-shaped or U-shaped, falling off gradually or suddenly, etc. A complete description of a frequency distribution is obtained when we have succeeded in writing the equation of the curve which "fits" the distribution. The type of curve to be fitted is chosen on the basis of certain criteria which are derived from the moments and the interrelations among the moments. The late Professor Karl Pearson developed the mathematics of a system of frequency curves and classified

distributions according to several "types" of curves, but a complete exposition of these types is beyond the scope of this text.

**Normal curve.** A bell-shaped curve which is often approximated closely by frequency distributions and which is intimately involved in much of statistical inference is known as the *normal curve*. We need to know in detail the properties of this curve.

The general equation of the normal curve can be written as

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{(X-M)^2}{2\sigma^2}} \quad (14)$$

in which  $y$  represents the height for any value of the variable  $x$ ,  $N$  is the number of cases,  $\sigma$  is the standard deviation,  $M$  is the mean of the distribution, and  $\pi$  (3.1416) and  $e$  (2.7183) are well-known mathematical constants. In order to write the equation of a particular normal curve, i.e., one which corresponds to a particular distribution, we need to know  $N$ ,  $M$ , and  $\sigma$ . This is the basis for the fact that, when we have the usual bell-shaped distribution, we need only the mean and standard deviation to describe it adequately. But in order to say that a given distribution is really normal, it is necessary to show that the  $g$ 's (as defined on p. 28) are zero or approximately zero.

Referring again to equation (14), we note that the numerator part of the exponent could be written in terms of deviation units, i.e., with  $x$  instead of  $X - M$ . The  $y$  for a positive deviation of, say, 10 will be exactly the same as for a negative 10 for the simple reason that the deviation value in the formula is squared. This indicates that the normal curve is symmetrical about the mean, and hence the mean and median coincide. When  $x = 0$ , i.e., when we take  $X = M$ ,  $y$  has its maximal value, and therefore the mean and mode coincide. For values of  $x$  other than zero, the height of the curve will be less. This is evident when we consider the fact that the exponent in equation (14) is negative. The height of the curve as we go in either direction from the mean becomes less and less (see Fig. 7a). This dropping off is slow at first, then rapid, and then slow again. If we take the maximum value of  $y$  (i.e., at the mean) as unity, the ordinate at the point  $.5\sigma$  from the mean is about .883; at  $1\sigma$ , about .606; at  $2\sigma$ , .135; and at  $3\sigma$ , .011. As we go still farther from the mean, the value of  $y$  becomes smaller, and as  $x$  approaches infinity,  $y$  approaches zero (asymptotic).

Theoretically, the curve never touches the base line, but so far as empirical distributions are concerned,  $y$  does become zero.

For both the frequency polygon and the histogram, the frequency for a given interval is represented along the  $y$  axis or ordinate, but for smoothed curves and for mathematical curves such as that defined by equation (14), it is advantageous to regard the area under the curve for a particular grouping interval on the  $x$  axis as indicating the frequency for that interval. Accordingly the total

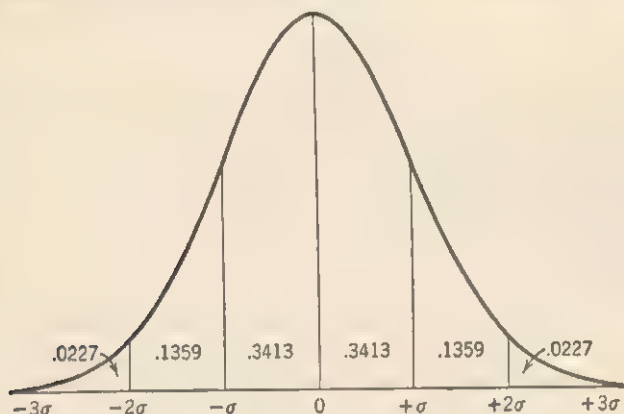


Fig. 7a. Normal curve.

area under the curve corresponds to the total frequency, or  $N$ , and the area under any given part of the curve, i.e., the area between any two  $x$  values, can be expressed as a percentage of the total. For example, the area included between the mean and the point on the base line  $1\sigma$  above the mean is 34.13 per cent of the total, and the area between plus and minus  $1\sigma$  is 68.26 per cent. The latter percentage has already been given on p. 26 as one way of interpreting the standard deviation. The limits plus and minus  $2\sigma$  will include 95.45 per cent; plus and minus  $3\sigma$ , 99.73 per cent; and plus and minus  $4\sigma$ , 99.9936 per cent. Theoretically, one must pass to plus and minus infinity to include all the area, but in practice 100 per cent of the cases will usually fall within the limits  $\pm 3\sigma$ , and nearly always within the limits  $\pm 4\sigma$ .

When we transform a set of scores to the so-called *standard score* form

$$z = \frac{X - M}{\sigma} = \frac{x}{\sigma} \quad (15)$$

we have each score expressed as a deviation from the mean in terms of multiples of the standard deviation of the original distribution. It can easily be shown that the standard deviation of our new set of scores will be unity, and the mean zero. The frequency polygon for the standard scores will have exactly the same shape as that for the original scores; this transformation is equivalent to translating the origin along the  $x$  axis to the point corresponding to the mean and changing the scale on the  $x$  axis so as to make the standard deviation equal to unity. If we let the total frequency be unity, we can think of the total area under the curve as being unity. This is equivalent to saying that  $N$  equals 1, and since with standard scores  $\sigma$  also equals 1, equation (14) can be written as

$$y = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad (16)$$

The value of  $1/\sqrt{2\pi}$  is about .39894, and therefore at  $z = 0$  (i.e., at the mean)  $y$  will equal .39894, which is the maximum  $y$  for the normal curve of unit area and unit standard deviation. The ordinates for other values of  $z$  will be less. For instance, at  $\pm 1z$ ,  $y = .24197$ , and at  $\pm 2z$ ,  $y = .05399$ .

The percentage area under any part of the curve can be determined by methods of the calculus. The area under the curve between any two values,  $z_1$  and  $z_2$ , is obtained as the value of the integral

$$A = \int_{z_1}^{z_2} y \, dz \quad (17)$$

Perhaps this expression will be more meaningful to the student who has not studied integral calculus if the given area is regarded as composed of a large number of strips, each having a tiny base  $dz$  and a height of  $y$ . For each such strip the area will be nearly  $ydz$ , and the integral sign in formula (17) simply means the "sum of" the areas of these tiny strips.

The student of the calculus will also note that the first derivative of either equation (14) or (16) set equal to zero and solved will yield a maximum for the curve when  $x$  or  $z$  equals zero, thus proving more rigorously that the mean and mode coincide. If the second derivative is set equal to zero and solved for  $x$  or  $z$ , it will

be found that the points of inflection of the curve are located where  $x$  is  $\pm\sigma$  or  $z$  is  $\pm 1$ .

**Normal curve table.** Because of the widespread use of the normal curve, tables of proportionate frequencies and ordinates for various  $z$  or  $x/\sigma$  values are available. The student need not be able to integrate equation (17) in order to understand a table of the normal curve functions. Table A of the Appendix contains four columns, the first of which is  $z$  or  $x/\sigma$  values. The second

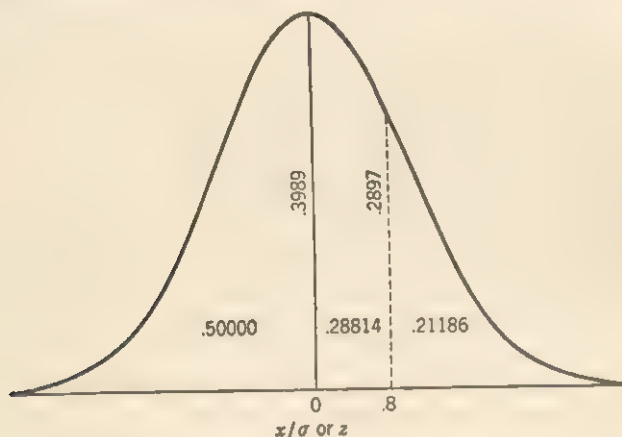


Fig. 7b. Normal curve functions.

column gives the area of the curve from the mean out to the corresponding  $z$  value, this area being the same whether  $z$  is positive or negative; a given  $z$  divides the curve into two parts, and the third column gives the area of the smaller part. The area of the larger part can be obtained by adding .5 to the entries in column 2. If one wishes to determine the proportionate area between plus and minus a given  $z$ , the values in column 2 should be doubled. The fourth column gives the  $y$  or ordinate for each of the  $z$  values. For purposes of reference, the meanings of the several entries in Table A are illustrated in Fig. 7b, in which an ordinate (dotted) has been erected at an  $x/\sigma$  value of  $+.8$ . The area from the mean to  $+.8$  is found from column 2 as .28814; the area below this point is .78814, and that above is .21186, of the total area. Note that .78814 plus .21186 equals unity and that .78814 is .50000 plus



.28814. The height of the curve at  $z = .8$  is found from column 4 as .2897, whereas the maximum height of .3989 is at the mean.

It is frequently useful to know the relationship between the various measures of dispersion for a normal distribution. It can be shown that the following hold true:

$$\begin{aligned} Q &= .8453 \quad AD = .6745 \quad SD \\ AD &= 1.1829 \quad Q = .7979 \quad SD \\ SD &= 1.4826 \quad Q = 1.2533 \quad AD \end{aligned}$$

It is also useful to know that for an  $N$  of 50 the  $SD$  will be about one-fifth the range, that for an  $N$  of 200 the  $SD$  will be about one-sixth the range, and that for an  $N$  of 1000 the  $SD$  will be about one-seventh the range.

The tabled values for the normal curve are often used in connection with problems similar to the following: If a distribution of the heights of men is normal with a mean of 68.0 inches and a standard deviation of 2.5, what percentage of men are more than 6 feet tall? We find  $z$  as the difference between 72 and 68, divided by  $\sigma$ , or  $z = 1.6$ ; then from Table A we find the percentage of cases which fall above this  $z$  value to be 5.48. Suppose that the mean IQ of 10-year-old boys is 100 and the standard deviation 16. What percentage have IQ's between 90 and 110? What percentage of 10-year-old boys would be classified as "gifted" (IQ above 140)?

The student will have noted that the answers to problems similar to the foregoing are possible by virtue of the fact that the areas and ordinates of Table A are for the standard score form of the normal curve with total area set equal to unity. By formula (15) one can pass from raw scores to standard scores and vice versa, and knowing  $N$  one can readily convert proportionate areas to frequencies or frequencies to proportions. Thus the table can be used with any normal distribution regardless of the original measurement units.

**Standard scores.** Perhaps it should be pointed out at this place that transforming scores, when distributions are normal or approximately so, to standard scores leads to new sets of scores which are comparable. For example, inches and pounds are not comparable units. If a man is 71 inches in height and weighs 170 pounds, it is impossible to say whether he is taller than he is heavy, but when the 71 inches is transformed to a  $z$  of .9 and the 170 pounds to a  $z$  of

1.3, we are able to say that, relative to his position in the two distributions, he is heavier than he is tall. Likewise, the raw scores on two psychological tests will seldom be comparable; changing to standard scores permits comparison, so that one can decide whether a boy's performance on one test is better or worse than his performance on another. This assumes, of course, a close approximation to normality, and that the means and standard deviations used in the transformations are based on the same or highly similar groups.

Standard scores, as defined by formula (15), will involve both positive and negative values and decimal scores. Since these are awkward to use, a further transformation is frequently made in such a way as to yield a distribution with a preassigned  $M$  and  $\sigma$ , instead of the  $M$  of 0 and  $\sigma$  of 1 which hold for the standard scores defined by formula (15). If we wish a distribution with a mean of 50 and a  $\sigma$  of 10, we can simply multiply each  $z$  by 10 and add 50. Multiplying each  $z$  by 20 and adding 100 would yield a mean of 100 and a  $\sigma$  of 20. Either of these transformations will get rid of negative values and permit a sufficient number of score values without the use of decimals. In general, if we wish to transform a set of scores having a mean,  $M$ , and a standard deviation,  $\sigma$ , to new values to be called  $Z$ 's, with mean equal to any value  $K$  and  $\sigma$  equal to  $S$ , all we need to do is to apply the relationship

$$Z = z(S) + K, \quad \text{or} \quad Z = \left( \frac{X - M}{\sigma} \right) (S) + K$$

which becomes

$$Z = \frac{S}{\sigma} (X) - \frac{M}{\sigma} (S) + K$$

The last form is the easier to use in practice, particularly with a calculating machine. Note that the last two terms will combine numerically and therefore can be placed in the lower dial as a positive or negative number; then the numerical value of  $S/\sigma$  can be set in the keyboard as a constant to be multiplied in turn upon the varying values of  $X$ . If the machine has a continuous upper dial, the best procedure is to multiply by the highest  $X$  first, and then, without clearing the dials, to subtract once for each successively lower value of  $X$ . Care is needed in aligning decimals, a check on which can be obtained by multiplying by the  $X$  nearest  $M$ . This

should lead to a value, in the lower dial, which is near  $K$ . With this setup, one can readily run off a table giving the values of  $Z$  for varying values of  $X$ .

The comparability of two sets of standard scores, either as  $z$ 's or as  $Z$ 's with the same mean ( $K$ ) and same  $\sigma$  ( $S$ ), does not hold for skewed distributions unless the two distributions show the same degree and direction of skewness. This is unlikely to be the case in practice. There is a scheme for use with skewed distributions which not only leads to comparable units but which also normalizes the distributions, i.e., changes the distributions from skewed to normal. This procedure is known as *T scaling*, and the resulting scores are known as *T scores*. They are usually so calculated as to yield a mean of 50 and a  $\sigma$  of 10, but other values for these constants are possible. The detailed procedure may be found in McCall's *Measurement*,\* which also includes a table for expediting the transformation. Suffice it to say here that *T scaling* basically involves determining the proportion (or percentage) of cases exceeding a given value plus half those reaching that value, and then entering such proportions in a table of the normal curve function to find the corresponding  $z$  values. Standard scores based on a normal distribution of original scores and *T scores* based on any shape distribution are comparable, provided they have been so determined as to yield the same mean and standard deviation. They differ only in the way in which they are computed, the standard score being a linear transformation which leaves the shape of the distribution unchanged, whereas *T scaling* changes the distribution to the normal form. If we begin with an exactly normal distribution and convert the scores to both  $z$ 's and  $T$ 's, there will be a linear correspondence between the two sets of transformed scores. If their means and sigmas are set equal, the  $Z$ 's and  $T$ 's will be equal to each other.

It will be recalled that the use of percentiles is another way of expressing scores on different tests so as to have comparability. The student should give sufficient thought to percentiles and standard scores to see how they are interrelated when the original scores are normal in distribution. Hint: The tabled functions (Table A) of the normal curve may help. The student might also demonstrate to his own satisfaction that the difference between the

\* McCall, W. A., *Measurement*, New York: Macmillan, 1939, pp. 505-508.

50th and 60th percentile points is *not* apt to be equal to the difference between the 80th and 90th percentile points.

**Kinds of distributions.** In anticipation of topics to be discussed, it might be well to mention some possible ways of regarding frequency distributions. We can have an *observed*, or *sample*, distribution of scores for a group of  $N$  individuals; we can imagine a *population* distribution of scores for either a finite or for an infinite  $N$ ; and we can conceive of a distribution curve defined by a *mathematical* equation (or function). Because of chance factors (as yet undefined herein) we do not expect an observed sample distribution to be exactly like the distribution of the population from which the sample is drawn or like a defined mathematical distribution.

Since we are seldom able to measure all members of a population, we can only assume that population scores follow some defined mathematical distribution. The form of mathematical curve assumed is usually decided upon by a consideration of the shape of an observed sample distribution. As will be seen later, the reasonableness of the assumption can be checked statistically.

It is possible, however, to show mathematically that under prescribed conditions given measures will follow a defined distribution curve exactly. We shall refer to such a distribution as *theoretical* or *expected*. Strictly speaking a mathematical distribution curve holds only for a continuous variable. If we had the distribution for a discrete variable, such as number of children per family, we would never expect that increasing  $N$  would produce a curve—the variable takes on only *point* values 0, 1, 2, etc.; hence we cannot allow the interval size (see p. 32) to approach zero, which is necessary for a smooth curve.

As implied above, there are distribution curves which are not normal. We shall introduce other curves (or functions) when needed. Thus far, the normal curve has been discussed as a *frequency* curve, and the area interpretation has been in terms of the number of individuals or percentage of cases falling between certain score limits. This same curve is often spoken of as the normal *probability* curve, and as such it is regarded as a theoretical curve. We shall see, moreover, that there are theoretical curves other than the normal curve which may be regarded as probability curves.

## Probability and Hypothesis Testing

Statistical inference and the testing of hypotheses involve the concept of chance, or probability. A simple example will serve to illustrate the probabilistic nature of hypothesis testing. Suppose a chap claims that he can distinguish between Camels and Lucky Strikes. To test his claim we could blindfold him and present him with either a Camel or a Lucky Strike (the brand to be presented is determined by tossing a coin). If on this 1 trial he correctly named the brand we would not be inclined to accept his claim since he would have a 50 50 chance of being correct on a sheer guessing basis. So we give him a second trial (again, and for any subsequent trials, we toss a coin to determine which brand to present to him). If he were again successful we might give some credence to his claim but someone might ask whether making 2 correct discriminations could happen on the basis of chance. We shall presently see that the chances are 1 in 4 of getting 2 correct, i.e., success on 2 trials could easily occur on the basis of chance.

But suppose he is correct on 3 trials, then on the fourth trial, and also on the fifth; or perhaps he is correct on 10 trials, or perhaps on 9 of 10 trials? Regardless of the number of trials and the number of successes we certainly should have some information about chance success, or the probability of correctly naming the brands on the basis of chance guessing, before we reach a decision regarding the claimed ability to distinguish between the 2 brands of cigarettes. This and similar decision problems involve notions of probability, to which we now turn.



**Probability.** If one had a box containing 70 white and 30 black balls, well mixed, and were to draw 1 ball at random, the chance of the drawn ball's being black is said to be 30 out of 100, and the chance of its being white would be .70. This can be interpreted to mean that, if we made 1000 random draws, each time replacing the drawn ball and remixing the contents of the box, the percentage of black balls drawn would be about 30, and of white draws about 70. If one rolls a die, the probability of obtaining a 4 is  $\frac{1}{6}$ ; i.e., a large number of rolls would yield a 4 about  $\frac{1}{6}$  of the time. If one tosses a symmetrical coin, it is usually said that there is a fifty-fifty chance of its landing "heads up," or the probability of a head is  $\frac{1}{2}$ . This is another way of saying that in the long run the proportion of times that the coin lands as a head will be the same as the proportion of times it lands as a tail.

These very simple examples illustrate a *definition of probability*: if an event can happen in  $A$  ways and fail in  $B$  ways, all possible ways being equally likely, the probability of its occurring is  $A/(A + B)$  and of its failing is  $B/(A + B)$ . That is, a probability figure is the ratio of the number of favorable events to the total number of events, and it is therefore necessary that we be able to enumerate events in order to arrive at a probability figure.

If we draw a card from a pack, the probability of obtaining a spade is  $\frac{1}{4}$ , and the probability of drawing a club is also  $\frac{1}{4}$ , but the probability of drawing *either* a spade *or* a club is  $\frac{1}{4}$  plus  $\frac{1}{4}$ , or  $\frac{1}{2}$ . If we roll a die, the probability of obtaining *either* a 4 *or* a 5 is  $\frac{1}{6}$  plus  $\frac{1}{6}$ , or  $\frac{1}{3}$ . These two situations illustrate the *addition theorem* of probability: the probability that *either* one event *or* another event will happen is the sum of the probabilities of their occurrences as single events. (The events must be mutually exclusive; i.e., if one occurs, the other cannot.)

If we roll a pair of dice, the probability of a 2 on the first *and* a 5 on the second is  $\frac{1}{6}$  times  $\frac{1}{6}$ , or  $\frac{1}{36}$ . If we toss 2 coins, the probability that the first will land a head *and* the second a head is  $\frac{1}{2}$  times  $\frac{1}{2}$ , or  $\frac{1}{4}$ , which is, of course, the probability that both will land as heads. Notice that the result obtained with the second die or coin is independent of the outcome of the first die or coin. These two examples illustrate the *multiplication theorem*: the probability of 2 (or more) independent events' occurring simultaneously or in succession (one *and* the other) is the product of their separate probabilities.

As just indicated, if one tosses 2 coins, the probability that the first will land a head *and* also the second a head will be  $\frac{1}{2}$  times  $\frac{1}{2}$ , or  $\frac{1}{4}$ , which is the probability that both will fall as heads. The probability that the first will land a head and the second a tail will also be  $\frac{1}{2}$  times  $\frac{1}{2}$ , or  $\frac{1}{4}$ . But 1 head and 1 tail can be obtained in a manner mutually exclusive to the above; i.e., the first can land as a tail and the second as a head, and this combination or event has a probability of  $\frac{1}{4}$ , whence the probability of obtaining 1 head and 1 tail will be  $\frac{1}{4}$  plus  $\frac{1}{4}$ , or  $\frac{1}{2}$ . This same result can be arrived at by listing all the possible combinations and taking the ratio of the number of favorable to the total number of possible combinations. The possible combinations are *HH, HT, TH, TT*, from which we see that 2 out of the 4 possible events are favorable for the occurrence of 1 head and 1 tail. We also note that 1 out of 4 is favorable to 2 heads.

Suppose we were to toss 3 coins; we would have the following possible combinations:

COIN 1	<i>H</i>	<i>H</i>	<i>H</i>	<i>H</i>	<i>T</i>	<i>T</i>	<i>T</i>	<i>T</i>
COIN 2	<i>H</i>	<i>H</i>	<i>T</i>	<i>T</i>	<i>H</i>	<i>H</i>	<i>T</i>	<i>T</i>
COIN 3	<i>H</i>	<i>T</i>	<i>H</i>	<i>T</i>	<i>H</i>	<i>T</i>	<i>H</i>	<i>T</i>

The total number of possible "events" is 8, 1 of which is favorable to 3 heads, 3 to 2 heads, 3 to 1 head, and 1 to no heads, thus giving the respective probabilities of  $\frac{1}{8}$ ,  $\frac{3}{8}$ ,  $\frac{3}{8}$ , and  $\frac{1}{8}$ . If we were to toss 4 coins, we would have the following probabilities:

4 heads	$\frac{1}{16}$	1 head	$\frac{4}{16}$
3 heads	$\frac{4}{16}$	0 head	$\frac{1}{16}$
2 heads	$\frac{6}{16}$		

The student should satisfy himself that these are the correct figures by writing down all the combinations possible and counting those favorable to any particular number of heads.

**Binomial distribution.** The process of determining possible combinations becomes quite laborious for, say, 10 coins, but the several probabilities can be obtained by the coefficients in the expansion of the binomial  $(a + b)^n$ . Thus for  $n = 2$  (i.e., 2 coins) we have  $a^2 + 2ab + b^2$ , or 1, 2, 1; for  $n = 3$ ,  $a^3 + 3a^2b + 3ab^2 + b^3$ , or 1, 3, 3, 1; for  $n = 4$  the coefficients are 1, 4, 6, 4, 1. In each case the sum of the coefficients,  $2^n$ , will be the total possible combinations, and the coefficients taken as ratios with the common

denominator,  $2^n$ , will represent the probabilities for  $n$ ,  $n - 1$ ,  $n - 2$ ,  $\dots$  0 heads.

The student may recall that the general expansion of  $(a + b)^n$  is

$$a^n + na^{n-1}b + \frac{n(n-1)}{1 \times 2} a^{n-2}b^2 + \frac{n(n-1)(n-2)}{1 \times 2 \times 3} a^{n-3}b^3 + \dots$$

This expansion will contain  $(n + 1)$  terms and will terminate in  $b^n$ . For  $n = 10$ , we have the following coefficients: 1, 10, 45, 120, 210, 252, 210, 120, 45, 10, 1, which sum to 1024, or 2 to the tenth power. Thus the probability that all 10 coins will fall as heads is  $1/1024$ ; 9 heads,  $10/1024$ ; etc. If we plot these values as a frequency polygon—these coefficients are frequencies in the sense that they represent the expected number of times for 10 heads, 9 heads, etc., out of a total of 1024 tosses—we will have a bell-shaped graph which will resemble somewhat the normal curve.

Another and more useful way, for our purpose, of considering the binomial expansion is to use  $p$  and  $q$ , in the place of  $a$  and  $b$ , with  $p$  defined as the probability of success on a single element and  $q$  as the probability of failure, or  $q = 1 - p$ . Thus we would have  $(p + q)^n$ . Suppose  $n = 2$ ; the expression would be  $p^2 + 2pq + q^2$ . If  $p = \frac{1}{2}$ , as in the coin situation, this would give  $(\frac{1}{2})^2 + 2(\frac{1}{2})(\frac{1}{2}) + (\frac{1}{2})^2$ , or  $\frac{1}{4}$ ,  $\frac{2}{4}$ , and  $\frac{1}{4}$  as the probabilities for securing 2 heads, 1 head, and 0 head respectively. Each term is itself a probability fraction; the numerators are 1, 2, and 1 as before. For  $n = 10$ , we would have  $(\frac{1}{2})^{10}$  or  $1/1024$ ,  $10(\frac{1}{2})^9(\frac{1}{2})$  or  $10/1024$ ,  $45(\frac{1}{2})^8(\frac{1}{2})^2$  or  $45/1024$ , etc., as the probabilities for obtaining 10 heads, 9 heads, 8 heads, etc.

The chief advantage of using the  $p$  and  $q$  notation is that we can readily see what happens when  $p$  is not equal to  $\frac{1}{2}$ . Consider the expectation when we roll a pair of dice with "success" defined as the rolling of "snake eyes." We would have  $(p + q)^2 = (\frac{1}{6} + \frac{5}{6})^2 = \frac{1}{36} + 2(\frac{5}{36}) + \frac{25}{36}$  as indicating the probability of obtaining 2 one-spots, 1 one-spot, and 0 one-spot. If 3 dice were rolled, we would have  $\frac{1}{216} + 3(\frac{5}{216}) + 3(\frac{25}{216}) + \frac{125}{216}$  or  $\frac{1}{216}$ ,  $\frac{15}{216}$ ,  $\frac{75}{216}$ , and  $\frac{125}{216}$  as the respective probabilities for 3, 2, 1, and 0 one-spots. The important thing for the student to note is that these probabilities are definitely skewed—not all probability distributions are of the symmetrical type. The student can, as a tedious exercise, work out the probabilities for 4, 5, 6, 7, and 8 dice, and therefrom learn

that the shape of the distribution changes from marked skewness to less and less skewness as the number of dice is increased. It can be easily shown that, if  $p = \frac{5}{6}$  and  $q = \frac{1}{6}$ , the skewness will be in the opposite direction. Another proposition which the student can demonstrate to himself is that, for a fixed  $n$ , the skewness increases as  $p$  is taken farther from  $\frac{1}{2}$  in either direction — extremely small or extremely large  $p$ 's (near unity) lead to very marked skewness.

The binomial expansion provides the probabilities of the theoretically expected frequencies for given  $n$ 's,  $p$ 's, and  $q$ 's. Such theoretical distributions can be described as to central value, variation, skewness, and kurtosis. The numerical values for these measures may be obtained by direct computation from the distributions built up by the binomial expansion, or these measures may be obtained by simple formulas, which can be derived by simple algebra, without having the actual distributions available.

The student can, as an exercise, perform an empirical check on the formulas for the mean and standard deviation. The formulas are:

$$M = np$$

$$\sigma = \sqrt{npq}$$

$$g_1 = \frac{q - p}{\sqrt{npq}} \quad (\text{skewness})$$

$$g_2 = \frac{1 - 6pq}{npq} \quad (\text{kurtosis})$$

It should be noted that  $n$  is the number of elements, not the number of cases. The formula for skewness permits several deductions. When  $p = \frac{1}{2}$ ,  $q$  also equals  $\frac{1}{2}$ , and hence the skewness is zero; the degree of skewness for a fixed  $n$  depends upon the deviation of  $p$  from  $\frac{1}{2}$ , i.e., the smaller or the larger the probability of success for each element, the more skewed the distribution. Note also that, since  $n$  is in the denominator, the larger the number ( $n$ ) of elements, the smaller the skewness for fixed values of  $p$  and  $q$ .

The above formulas describe the theoretically expected distribution for given  $n$ 's,  $p$ 's, and  $q$ 's. As will be seen later, any empirical distribution obtained by tossing 10 coins or rolling 3 dice will

1997

1998

1999

2000

2001

2002

2003

2004

2005

2006

2007

2008

2009

2010

2011

2012

2013

2014

2015

2016

2017

2018

2019

2020

2021

2022

2023

2024

2025

2026

2027

2028

2029

2030

2031

2032

2033

2034

2035

2036

2037

2038

2039

2040

2041

2042

2043

2044

2045

2046

2047

2048

2049

2050

2051

2052

2053

2054

2055

2056

2057

2058

2059

2060

2061

2062

2063

2064

2065

2066

2067

2068

2069

2070

2071

2072

2073

2074

2075

2076

2077

2078

2079

2080

2081

2082

2083

2084

2085

2086

2087

2088

2089

2090

2091

2092

2093

2094

2095

2096

2097

2098

2099

2100

2101

2102

2103

2104

2105

2106

2107

2108

2109

2110

2111

2112

2113

2114

2115

2116

2117

2118

2119

2120

2121

2122

2123

2124

2125

2126

2127

2128

2129

2130

2131

2132

2133

2134

2135

2136

2137

2138

2139

2140

2141

2142

2143

2144

2145

2146

2147

2148

2149

2150

2151

2152

2153

2154

2155

2156

2157

2158

2159

2160

2161

2162

2163

2164

2165

2166

2167

2168

2169

2170

2171

2172

2173

2174

2175

2176

2177

2178

2179

2180

2181

2182

2183

2184

2185

2186

2187

2188

2189

2190

2191

2192

2193

2194

2195

2196

2197

2198

2199

2200

2201

2202

2203

2204

2205

2206

2207

2208

2209

2210

2211

2212

2213

2214

2215

2216

2217

2218

2219

2220

2221

2222

2223

2224

2225

2226

2227

2228

2229

2230

2231

2232

2233

2234

2235

2236

2237

2238

2239

2240

2241

2242

2243

2244

2245

2246

2247

2248

2249

2250

2251

2252

2253

2254

2255

2256

2257

2258

2259

2260

2261

2262

2263

2264

2265

2266

2267

2268

2269

2270

2271

2272

2273

2274

2275

2276

2277

2278

2279

2280

2281

2282

2283

2284

2285

2286

2287

2288

2289

2290

2291

2292

2293

2294

2295

2296

2297

2298

2299

2300

2301

2302

2303

2304

2305

2306

2307

2308

2309

2310

2311

2312

2313

2314

2315

2316

2317

2318

2319

2320

2321

2322

2323

2324

2325

2326

2327

2328

2329

2330

2331

2332

2333

2334

2335

2336

2337

2338

2339

2340

2341

2342

2343

2344

2345

2346

2347

2348

2349

2350

2351

2352

2353

2354

2355

2356

2357

2358

2359

2360

2361

2362

2363

2364

2365

2366

2367

2368

2369

2370

2371

2372

2373

2374

2375

2376

2377

2378

2379

2380

2381

2382

2383

2384

2385

2386

2387

2388

2389

2390

2391

2392

2393

2394

2395

2396

2397

2398

2399

2400

2401

2402

2403

2404

2405

2406

2407

2408

2409

2410

2411

2412

2413

2414

2415

2416

2417

2418

2419

2420

2421

2422

2423

2424

2425

2426

2427

2428

2429

2430

2431

2432

2433

2434

2435

2436

2437

2438

2439

2440

2441

2442

2443

2444

2445

2446

2447

2448

2449

2450

2451

On arrival of the vessel at the wharf

[illegible]

Abstract: This is a descriptive study of the experiences of 100 young women in the United States who have been sexually abused. The study was conducted in 1998 and 1999. The results show that the majority of the women were abused by a family member or a friend. The study also found that the majority of the women were abused at a young age. The study has implications for the development of interventions to help young women who have been sexually abused.

\_\_\_\_\_

[illegible]

Period	1900	1905	1910	1915
1	0	0	0	0.01
2	0	0	0	0.02
3	0	0.01	0	0.03
4	0	0.02	0	0.04
5	0	0.03	0	0.05
6	0.01	0.04	0.01	0.06
7	0.02	0.05	0.02	0.07
8	0.03	0.06	0.03	0.08
9	0.04	0.07	0.04	0.09
10	0.05	0.08	0.05	0.10
11	0.06	0.09	0.06	0.11
12	0.07	0.10	0.07	0.12
13	0.08	0.11	0.08	0.13
14	0.09	0.12	0.09	0.14
15	0.10	0.13	0.10	0.15
16	0.11	0.14	0.11	0.16
17	0.12	0.15	0.12	0.17
18	0.13	0.16	0.13	0.18
19	0.14	0.17	0.14	0.19
20	0.15	0.18	0.15	0.20
21	0.16	0.19	0.16	0.21
22	0.17	0.20	0.17	0.22
23	0.18	0.21	0.18	0.23
24	0.19	0.22	0.19	0.24
25	0.20	0.23	0.20	0.25
26	0.21	0.24	0.21	0.26
27	0.22	0.25	0.22	0.27
28	0.23	0.26	0.23	0.28
29	0.24	0.27	0.24	0.29
30	0.25	0.28	0.25	0.30
31	0.26	0.29	0.26	0.31
32	0.27	0.30	0.27	0.32
33	0.28	0.31	0.28	0.33
34	0.29	0.32	0.29	0.34
35	0.30	0.33	0.30	0.35
36	0.31	0.34	0.31	0.36
37	0.32	0.35	0.32	0.37
38	0.33	0.36	0.33	0.38
39	0.34	0.37	0.34	0.39
40	0.35	0.38	0.35	0.40
41	0.36	0.39	0.36	0.41
42	0.37	0.40	0.37	0.42
43	0.38	0.41	0.38	0.43
44	0.39	0.42	0.39	0.44
45	0.40	0.43	0.40	0.45
46	0.41	0.44	0.41	0.46
47	0.42	0.45	0.42	0.47
48	0.43	0.46	0.43	0.48
49	0.44	0.47	0.44	0.49
50	0.45	0.48	0.45	0.50
51	0.46	0.49	0.46	0.51
52	0.47	0.50	0.47	0.52
53	0.48	0.51	0.48	0.53
54	0.49	0.52	0.49	0.54
55	0.50	0.53	0.50	0.55
56	0.51	0.54	0.51	0.56
57	0.52	0.55	0.52	0.57
58	0.53	0.56	0.53	0.58
59	0.54	0.57	0.54	0.59
60	0.55	0.58	0.55	0.60
61	0.56	0.59	0.56	0.61
62	0.57	0.60	0.57	0.62
63	0.58	0.61	0.58	0.63
64	0.59	0.62	0.59	0.64
65	0.60	0.63	0.60	0.65
66	0.61	0.64	0.61	0.66
67	0.62	0.65	0.62	0.67
68	0.63	0.66	0.63	0.68
69	0.64	0.67	0.64	0.69
70	0.65	0.68	0.65	0.70
71	0.66	0.69	0.66	0.71
72	0.67	0.70	0.67	0.72
73	0.68	0.71	0.68	0.73
74	0.69	0.72	0.69	0.74
75	0.70	0.73	0.70	0.75
76	0.71	0.74	0.71	0.76
77	0.72	0.75	0.72	0.77
78	0.73	0.76	0.73	0.78
79	0.74	0.77	0.74	0.79
80	0.75	0.78	0.75	0.80
81	0.76	0.79	0.76	0.81
82	0.77	0.80	0.77	0.82
83	0.78	0.81	0.78	0.83
84	0.79	0.82	0.79	0.84
85	0.80	0.83	0.80	0.85
86	0.81	0.84	0.81	0.86
87	0.82	0.85	0.82	0.87
88	0.83	0.86	0.83	0.88
89	0.84	0.87	0.84	0.89
90	0.85	0.88	0.85	0.90
91	0.86	0.89	0.86	0.91
92	0.87	0.90	0.87	0.92
93	0.88	0.91	0.88	0.93
94	0.89	0.92	0.89	0.94
95	0.90	0.93	0.90	0.95
96	0.91	0.94	0.91	0.96
97	0.92	0.95	0.92	0.97
98	0.93	0.96	0.93	0.98
99	0.94	0.97	0.94	0.99
100	0.95	0.98	0.95	1.00

The following table gives the results of the tests of the hypothesis of normality of the distribution of the data.

Period	1900	1905	1910	1915
1	0.01	0.01	0.01	0.01
2	0.02	0.02	0.02	0.02
3	0.03	0.03	0.03	0.03
4	0.04	0.04	0.04	0.04
5	0.05	0.05	0.05	0.05
6	0.06	0.06	0.06	0.06
7	0.07	0.07	0.07	0.07
8	0.08	0.08	0.08	0.08
9	0.09	0.09	0.09	0.09
10	0.10	0.10	0.10	0.10
11	0.11	0.11	0.11	0.11
12	0.12	0.12	0.12	0.12
13	0.13	0.13	0.13	0.13
14	0.14	0.14	0.14	0.14
15	0.15	0.15	0.15	0.15
16	0.16	0.16	0.16	0.16
17	0.17	0.17	0.17	0.17
18	0.18	0.18	0.18	0.18
19	0.19	0.19	0.19	0.19
20	0.20	0.20	0.20	0.20
21	0.21	0.21	0.21	0.21
22	0.22	0.22	0.22	0.22
23	0.23	0.23	0.23	0.23
24	0.24	0.24	0.24	0.24
25	0.25	0.25	0.25	0.25
26	0.26	0.26	0.26	0.26
27	0.27	0.27	0.27	0.27
28	0.28	0.28	0.28	0.28
29	0.29	0.29	0.29	0.29
30	0.30	0.30	0.30	0.30
31	0.31	0.31	0.31	0.31
32	0.32	0.32	0.32	0.32
33	0.33	0.33	0.33	0.33
34	0.34	0.34	0.34	0.34
35	0.35	0.35	0.35	0.35
36	0.36	0.36	0.36	0.36
37	0.37	0.37	0.37	0.37
38	0.38	0.38	0.38	0.38
39	0.39	0.39	0.39	0.39
40	0.40	0.40	0.40	0.40
41	0.41	0.41	0.41	0.41
42	0.42	0.42	0.42	0.42
43	0.43	0.43	0.43	0.43
44	0.44	0.44	0.44	0.44
45	0.45	0.45	0.45	0.45
46	0.46	0.46	0.46	0.46
47	0.47	0.47	0.47	0.47
48	0.48	0.48	0.48	0.48
49	0.49	0.49	0.49	0.49
50	0.50	0.50	0.50	0.50
51	0.51	0.51	0.51	0.51
52	0.52	0.52	0.52	0.52
53	0.53	0.53	0.53	0.53
54	0.54	0.54	0.54	0.54
55	0.55	0.55	0.55	0.55
56	0.56	0.56	0.56	0.56
57	0.57	0.57	0.57	0.57
58	0.58	0.58	0.58	0.58
59	0.59	0.59	0.59	0.59
60	0.60	0.60	0.60	0.60
61	0.61	0.61	0.61	0.61
62	0.62	0.62	0.62	0.62
63	0.63	0.63	0.63	0.63
64	0.64	0.64	0.64	0.64
65	0.65	0.65	0.65	0.65
66	0.66	0.66	0.66	0.66
67	0.67	0.67	0.67	0.67
68	0.68	0.68	0.68	0.68
69	0.69	0.69	0.69	0.69
70	0.70	0.70	0.70	0.70
71	0.71	0.71	0.71	0.71
72	0.72	0.72	0.72	0.72
73	0.73	0.73	0.73	0.73
74	0.74	0.74	0.74	0.74
75	0.75	0.75	0.75	0.75
76	0.76	0.76	0.76	0.76
77	0.77	0.77	0.77	0.77
78	0.78	0.78	0.78	0.78
79	0.79	0.79	0.79	0.79
80	0.80	0.80	0.80	0.80
81	0.81	0.81	0.81	0.81
82	0.82	0.82	0.82	0.82
83	0.83	0.83	0.83	0.83
84	0.84	0.84	0.84	0.84
85	0.85	0.85	0.85	0.85
86	0.86	0.86	0.86	0.86
87	0.87	0.87	0.87	0.87
88	0.88	0.88	0.88	0.88
89	0.89	0.89	0.89	0.89
90	0.90	0.90	0.90	0.90
91	0.91	0.91	0.91	0.91
92	0.92	0.92	0.92	0.92
93	0.93	0.93	0.93	0.93
94	0.94	0.94	0.94	0.94
95	0.95	0.95	0.95	0.95
96	0.96	0.96	0.96	0.96
97	0.97	0.97	0.97	0.97
98	0.98	0.98	0.98	0.98
99	0.99	0.99	0.99	0.99
100	1.00	1.00	1.00	1.00



The following table gives the results of the tests of the hypothesis of normality of the distribution of the data.



all the bars ( $= 65,536$ ) will give the probability value of .03841 reported above. To approximate this by the normal curve we need to consider the area under the curve for that part of the curve which spans the bars with base-line values of 12, 13, 14, 15, and 16. Obviously we need the area under the curve beyond an  $X$  value of 11.5, a value which doesn't make much sense in terms of number of heads but which does make sense when it is recalled that we are here treating a point (discrete) variable as though it were a continuous variable, normally distributed. Hence we have  $X - M = 11.5 - 8 = 3.5 = x$ , and  $x/\sigma = 3.5/2 = 1.75$ . Turning to Table A we find that the proportionate area under a normal curve beyond an  $x/\sigma$  of 1.75 is .04056. This is our approximation to the exact probability value of .03841; the error in this approximation is of the order of .002. In general, when  $n$  is fairly large the failure to shift .5 (e.g., from 12 to 11.5 as done here) leads to a negligible error. This shift of .5 is referred to as *correction for continuity*.

We can, of course, use the normal curve to approximate any of the exact probabilities obtainable from Table 5 (or from the binomial with  $n$  other than 16). For example, the exact probability of obtaining 10 or 11 or 12 heads is  $(8008 + 4368 + 1820)/65,536$ , or .21661. The normal curve approximation, calculated as the proportionate area under the curve from 9.5 to 12.5, is .21441.

It is fortunate for us that for  $n$  larger and larger the normal curve approximation becomes better and better since for  $n$  large the computation of exact probabilities by the binomial method becomes very arduous.

Notice that, in approximating the probability, we have utilized an *area* under a curve; i.e., we have said that the area between 2  $X$  values taken relative to a total area may be interpreted as a probability figure. This is not inconsistent with our original definition of probability involving number (frequency) of events favorable relative to a total number of events (total frequency). Since, as previously indicated, the total area under a frequency curve for a continuous variable (or function) can be regarded as the total frequency, and the area for a particular segment can be regarded as the frequency with which values (or scores) fall in the given segment, it follows that the ratio of the segmental to the total frequency may be spoken of as a probability—the probability that a score falls between the 2  $X$  values defining the segment. When we are dealing with a distribution of the normal type, the

probability associated with a given segment is found by converting the 2  $X$  values, which define an interval, into  $x/\sigma$  values and then determining the area from Table A. The obtained proportionate area represents the probability expressed as a decimal fraction.

It should be obvious that, when we consider the unit normal curve, we can readily specify the proportionate area between any 2  $x/\sigma$  values, say  $z_1$  and  $z_2$ , and interpret the proportion as the probability of obtaining  $x/\sigma$  values between the given  $z_1$  and  $z_2$ . By reference to tables more extensive than Table A, it can be found that the area between an  $x/\sigma$  of  $-1.96$  and an  $x/\sigma$  of  $+1.96$  is very nearly .95; hence it would be said that .95 represents the probability of obtaining  $x/\sigma$  values between these 2 points. Furthermore, it can be said that .05 represents the probability that an  $x/\sigma$ , drawn at random from a normally distributed supply of  $(x/\sigma)$ 's, will be numerically larger than 1.96. Similarly it can be said that the probability of drawing an  $x/\sigma$  between  $\pm 2.576$  is very nearly .99, while the probability for an  $x/\sigma$  falling outside these limits is .01.

The foregoing interpretation of proportionate areas under the normal curve as probabilities is, in a sense, the basis for sometimes calling this curve the *normal probability curve*. It has been noted that for  $p$  not equal to  $q$ , the point binomial leads to skewed probability distributions. For continuous functions it is also possible to have distributions, other than the normal, which permit probability statements on the basis of proportionate areas. Later we shall consider the use of 3 nonnormal probability distributions.

## HYPOTHESIS TESTING

We may now return to a consideration of the blindfold test of the claimed ability to distinguish between 2 cigarette brands. By using the binomial expansion we can readily specify the probability of being correct (by chance)  $n$  times out of  $n$  trials. The answer is simply  $1/2^n$ ; if there were 10 trials the probability of 10 correct choices (by chance guessing—no real discriminatory ability) would be  $1/1024$ , or about .001; the probability of being correct 16 out of 16 trials would be  $1/65,536$ , or about .000015. If our self-proclaimed expert did succeed in 10 of 10 trials we would, because of the small probability of 10 successes by chance, concede that he

really possessed the ability to discriminate between the 2 brands.

But suppose he was successful on 9 trials of a 10-trial series? We could readily specify the probability of 9 successes by chance (it would be  $10/1024$ ) but for reasons which will become apparent later, it is better to ascertain the probability of as many as 9 successes in 10 trials (at least 9, or 9 or more, successes). This probability will be the probability of exactly 9 successes plus the probability of exactly 10 successes, or  $10/1024 + 1/1024 = 11/1024 =$  about .01, which is sufficiently small that we might decide that his performance was based on ability rather than on chance. Note that such a record would occur by chance about 1 time in 100, so we couldn't be sure that he really had the ability.

Next, let us suppose that he was correct on 8 of the 10 trials. The probability of at least 8 successes occurring on a chance basis would be  $45/1024 + 10/1024 + 1/1024 = 56/1024 =$  about .05. Would we now conclude that he had the claimed ability? If we did so conclude we wouldn't be as sure of our inference as when there were 9 successes, and far less sure than when there were 10 successes. In other words, the smaller the probability of attaining an obtained number of successes by chance the surer we would be of our conclusion. If he were successful on 7 trials (probability =  $P = .17$  for 7 or more successes) we would no doubt hesitate before conceding that his performance was based on ability to discriminate, since 7 successes can too easily occur on the basis of chance alone.

We are thus led to the question: What level of probability should one adopt as a criterion for deciding whether an observed performance is based on ability rather than chance? We are not yet ready to attempt an answer to this, but it might be remarked here that in choosing a level of probability it is necessary to consider the risk of being wrong in concluding that the fellow can discriminate vs. the risk of attributing his performance to chance when in reality he does have some ability.

Whether a person can discriminate between 2 brands of cigarettes is a simple illustration of the problem of statistical inference, or the testing of hypotheses. For purpose of inference we set the hypothesis that our friend cannot discriminate between brands. This readily permits us to calculate the probability ( $P$ ) of as many successes by chance as he attains on a series of trials; if  $P$  is suffi-

ciently small we reject the hypothesis of no ability, and in so doing we are saying that his number of successes is *statistically significant*, that is, nonchance. The *level of significance* associated with rejection of the hypothesis is represented by a probability—if we agree to reject the hypothesis only when the probability of chance success is as low as .01, we will have adopted the  $P = .01$  level of significance. If we are willing to be less sure and require  $P$  to be as low as .05 we will be working at the .05 level of significance. Whether we adopt the .01 or the .05 level is somewhat arbitrary—for this chapter let us quite arbitrarily choose  $P = .01$  as our working level of significance. After considering the more detailed discussion of this issue later in the chapter, the reader may prefer to adopt the .05 or some other level for judging significance.

The binomial expansion (and normal curve approximation thereto) may be used in a wide variety of situations as a means of testing hypotheses. A general requisite is that we be able to specify the probability of success (or something analogous to success) for a single element (coin, die, trial, etc.). In other words, we need to specify  $p$  (and  $q$ ) so as to use  $(p + q)^n$  or we need to calculate the mean and  $SD$  in order to utilize the normal curve approximation when  $n$  is not small.

Consider the problem of public opinion polling. In polling studies we are usually interested in whether or not a population of potential voters is split 50-50 on an issue. Accordingly we set the hypothesis that there is a 50 50 split in the population. This hypothesis is to be accepted or rejected on the basis of information yielded by a sample of  $N$  persons, who are asked to respond "yes" (agree) or "no" (disagree) to a statement of the given issue. Suppose for sake of simplicity we take  $N = 64$  and that 42 of them give a yes response. Is this result consistent with the hypothesis of a 50-50 split?

To answer this we note that so far as the opinion poller is concerned there is, by hypothesis, a 50 50 chance that any individual in the sample will say yes (this despite the fact that the individual so far as *he* is concerned is not giving a chance response). Thus the probability of a yes response for a single individual is  $1/2$ ; that is,  $p = .5$  and  $q = .5$  (since  $q$  is always  $1 - p$ ). Now our sample of 64 is analogous to a trial toss of 64 coins, so we consider the binomial distribution with  $n = N = 64$ . The mean  $= Np = 32$ , and the  $SD = \sqrt{Npq} = 4$ . The number of yes responses, 42,

deviates 10 from the mean. Our normal curve approximation would be slightly better if we used  $41.5 = 32 + 9.5$  as our deviation correction for continuity. Thus we have  $z = 10.1 = 2.50$ . Looking to Table A we find that the probability of obtaining this large a deviation in a specified direction from 32 is about .006. But in testing our hypothesis of 30:30 split we need also to include the probability of obtaining as large a deviation in the opposite direction. Then we double .006 and have  $P = .012$  as the probability of a large deviation irrespective of direction. Since the  $P$  is very near our arbitrarily and temporarily agreed upon  $P = .01$  level for judging against, we reject the hypothesis of an equal split in the population being sampled, and this rejection implies that a majority of the population would endorse the given statement.

In passing it should be noted that had the number of yes or no votes been 45 to 45, we could accept the hypothesis of an equal split. But this acceptance would not prove the hypothesis true. It would only be a chance deviation from any of an infinite of splits, 45:45, 44:46, etc. Although we would have more to say later.

Our sample results are usually expressed in percentage form. Thus, in this example, we might say .500. If we the hypothesis of a 50:50 split in the population, 50% or 50 percent yeses and 50 percent of 42 would be predicted. If 60% or 60 percent the hypothesis would be. A sample of 50 or 100 would then in testing the right tail of a normal curve of 42 mean 42 one has been testing the deviation of 18 from 40 or a percentage deviation of 45% from 40 in percentage form.

Another statistic for assessing the data is

$$\frac{s}{\sigma} = \frac{42 - 32}{\sqrt{64(5)(5)}} = 2.50$$

In assessing the goodness of fit of the data generated from the test by  $df = 42 - 1 = 41$  degrees of freedom, the denominator of 64 out of 49 are well within the bounds of error. Hence we have

$$\frac{s}{\sigma} = \frac{42 - 32}{\sqrt{\frac{41}{42} \cdot 64(5)(5)}} = \frac{10.00}{3.99} = 2.51$$

which differs from 2.50 only because of rounding errors. This

implies that dividing by  $N$  somehow preserves the  $z$ - $\sigma$  nature of the result. The numerator of  $z$  is a deviation—the deviation of an observed sample proportion from a hypothetical proportion. One might therefore deduce that the denominator is a  $\sigma$ , but a  $\sigma$  of what?

Let  $x$  = number of yeses, frequency of yeses.  $x$  can vary from zero to  $N$  with  $M_x = Np$  and  $\sigma_x = \sqrt{Npq}$ . If we divide every possible  $x$  by  $N$  we have proportions. The mean in proportion units will be  $M_x/N = Np/N$  or simply  $p$  and by a principle hinted at on p. 26, the standard deviation in proportion units will be  $\sigma_x/N = \sqrt{Npq}/N = \sqrt{pq/N}$ . This last term is precisely what we have above as a denominator, hence as a  $\sigma$  it is the standard deviation of a distribution of proportions; we may symbolize this *SD* as  $\sigma_p$ .

In summary, we have  $np$  and  $\sqrt{npq}$  as the mean and *SD* for a chance distribution of successes on  $n$  coins or  $n$  dice or  $n$  trials, etc. We have  $Np$  and  $\sqrt{Npq}$  as the mean and *SD* of a chance distribution of number of yes responses for  $N$  individuals. We have  $p$  and  $\sqrt{pq/N}$  as the mean and *SD* of a chance distribution of proportion of yeses based on samples of  $N$  individuals. In the coin-tossing and analogous situations each toss or trial leads to a countable number of successes, and the distribution of the number of successes on successive trials follows the binomial. For the polling situation each sample of  $N$  cases leads to a *proportion* of yeses, and the distribution of proportions for successive samples of same size also tends to follow the binomial. Such a distribution is referred to as the *random chance sampling distribution of proportions*.

It is customary to refer to  $\sigma_p$  as the *standard error* of a proportion. The term *error* is used here because, in effect, we are measuring the variability due to chance sampling error. Actually, the sampling distribution of proportions is a theoretical distribution; we actually have just 1 sample proportion (or a few at most). Statistical theory provides us with information concerning the center, range, variability, and shape of the distribution to be expected if we did have a very large number of sample proportions.

The scheme outlined above for testing hypotheses is not of course restricted to the separate binomial test and the pooling situation. In the first place, the  $p$  for the binomial need not be 1/2—our setup might involve  $k/p$  of 1/3 (e.g., identifying 1 of 3 test are we confined to the hypothesis of 50/50 split when pooling



(e.g., we might be interested in whether there is a 2 to 1 split). In the second place, we need not limit ourselves to number of successes or number of yeses. The fundamental requirement is that we be able to categorize observations (or individuals) into 2 classes (a dichotomy) such as pass or fail, agree or disagree, like or dislike, present or absent, cured or not cured, etc.

When a hypothesis involving a proportion is tested, the general procedure is to express the observed proportion,  $p_{ob}$ , as a deviation from  $p_h$ , the proportion expected on the basis of a statistical hypothesis, then to divide this deviation by

$$\sigma_p = \sqrt{\frac{p_h q_h}{N}} \quad (18a)$$

This gives an  $x/\sigma$ , sometimes called a critical ratio (CR), which for  $N$  not too small and  $p_h$  not too extreme will follow the unit normal curve, the table of which permits us to ascertain the probability of a deviation as great as that observed. Note that the proviso that  $p_h$  can't be extreme follows from the fact that the binomial distribution is skewed when  $p$  is extreme, say when  $p$  is greater than .90 or less than .10 (see the formula for skewness on p. 45). Since the skewness is also a function of  $n$  it follows that any rule that we might adopt to prevent unjustifiable use of the normal curve approximation will be a function of  $N$  and  $p_h$ . In general when both  $Np_h$  and  $Nq_h$  exceed 5 we can safely use the normal curve; if either product is between 5 and 10 we should deduct  $.5/N$  from the *numerical* value of the deviation of  $p_{ob}$  from  $p_h$ . This is another way of incorporating the correction for continuity (p. 48).

Formula (18a) for  $\sigma_p$  has been written with  $p_h$  as a value specified by the hypothesis to be tested. As such the formula measures the chance variation in proportions when the hypothesis is true. Actually, saying "if there is a 50-50 split in opinion" is the same as saying "if the proportion of yeses is .50 in the population." If we let  $p_{pop}$  stand for population proportion then the variation of sample proportions is given by substituting  $p_{pop}$  (and  $q_{pop}$ ) in (18a). When one has an obtained proportion,  $p_{ob}$ , and doesn't know  $p_{pop}$  (usually the case) and has no hypothesis in mind, one uses  $p_{ob}$  as an estimate of  $p_{pop}$ , and

$$\sigma_p = \sqrt{\frac{p_{ob} q_{ob}}{N}} \quad (18b)$$

as an approximation of the standard error of an observed proportion.

At this point the student may be somewhat confused by the use of  $p$ , first as the probability of, say, success on a single element and then as a proportion. Note, however, that if we were told that .30 (a proportion) of a given group have brown eyes, we could say that the probability that a randomly selected person has brown eyes is .30. Furthermore, when we say that the probability of rolling a snake eye is  $1/6$  or .1667, we mean that the proportion of snake eyes for a large number of rolls will tend to be .1667.

**Some sampling theory.** To facilitate later discussion we shall now introduce some notions of sampling theory. We will confine our attention to what is known as simple *random sampling*. The conditions for random sampling are that each individual (person, plant, animal, observation, etc.) in a defined population (universe, or supply) shall have an equal chance of being included in the sample, and that the drawing of one individual shall in no way affect the drawing of another (that is, the drawings must be independent of each other). The first condition is not easily met in practice. The aim is, of course, to obtain a sample which will be, within limits of random or chance errors, representative of the population from which it is drawn.

When dealing with attributes, or the classification of individuals into 2 (or more) categories, for which the proportion in a given category is a useful descriptive measure, we can conceive of a population proportion,  $p_{pop}$ , and a proportion,  $p_{ob}$ , obtained on a random sample of  $N$  cases. Now if we could draw successive samples of  $N$ , determine  $p_{ob}$  for each sample, and then make a distribution of the several  $p_{ob}$  values, we would expect this distribution to follow the normal curve for  $N$  not small and  $p_{pop}$  not extreme. This follows from our discussion of the binomial distribution and normal curve approximation thereto, the only difference being that we were then speaking of a chance distribution about some hypothetical proportion,  $p_h$ . If  $p_h$  happened to equal  $p_{pop}$  we would be dealing with precisely the same distribution of sample values. If, for example, the hypothesis of a 50-50 split is true we would expect the distribution of successive sample proportions to center at .5 and have an  $SD = \sigma_p = \sqrt{p_h q_h / N} = \sqrt{(.5)(.5) / N}$ ; if the population proportion,  $p_{pop}$ , is .5 we would expect the successive sample proportions to have a mean of .5 and an  $SD = \sigma_p = \sqrt{p_{pop} q_{pop} / N} = \sqrt{(.5)(.5) / N}$ .

## DIFFERENCES BETWEEN PROPORTIONS

The testing of hypotheses need not be confined to a single proportion. This is fortunate because in research involving attributes we are more apt to have 2 proportions, and since each is subject to chance (sampling) error, it follows that the difference between them will also be subject to chance error. To test a hypothesis regarding the difference between 2 proportions it will be necessary that we have information concerning the theoretical random (chance) sampling distribution of the differences between proportions. We will need to distinguish 2 different types of situations: (1) proportions based on 2 samples drawn independently from 2 populations and (2) proportions for responses or observations obtained under 2 different conditions on just 1 sample. For either situation we set up a statistical hypothesis known as a *null hypothesis*. This hypothesis, which states that there is no difference between the population proportions, will be rejected if the obtained difference reaches some prescribed level of significance but will be accepted otherwise. Stated differently, if the observed difference could readily arise on a chance basis we accept the null hypothesis; if the probability of its occurrence by chance is small we reject the null hypothesis. Note that our statistical hypothesis of no difference may be, and often is, diametrically opposed to the research hypothesis being checked by the data. That is, on the basis of theory or prior observations we may expect a difference, yet for statistical reasons we set the null hypothesis. If the obtained difference is statistically significant in the expected direction we regard the data as tending to support the research hypothesis.

**Nonindependent proportions.** We shall consider first the situation in which the 2 proportions being compared are not based on independent groups but on just 1 group (or on 2 related groups). Suppose we are interested in whether a movie leads to a change of opinion, i.e., to an increase in the proportion favorable to some issue. We select a random sample from some defined population, get a yes (favorable) or no (unfavorable) response from each individual, show them the movie, then again get a yes or no response from each. Our next step is that of tabulation and, since we are concerned with possible changes in opinion, we will need to arrange our tabulation so as to show how many changed from no to yes, how many from yes to no, and how many "stood pat." This can be done by placing tally marks in a 2 by 2, or fourfold,

table such as that depicted in Table 6. For an individual who gave a yes response the first time and a yes response the second time, a tally would go in the upper right-hand cell; for a yes at first followed by a no, a tally would go in the upper left quadrant; and so on.

Table 6. TABULATION PLAN FOR HANDLING PROPORTIONS BASED ON THE SAME INDIVIDUALS

		Frequencies			Proportions				
		2nd			2nd				
		No	Yes		No	Yes			
1st	Yes	A	B	A + B	1st	a	b	p <sub>1</sub>	
	No	C	D	C + D		c	d	q <sub>1</sub>	
		A + C	B + D	N			q <sub>2</sub>	p <sub>2</sub>	1.0

Let  $A$ ,  $B$ ,  $C$ , and  $D$  represent the respective frequencies for yes-no, yes-yes, no-no, and no-yes responses. Then  $A + B$  is the total number of yeses at first and  $B + D$  is the total number of yeses the second time. If each of these totals is divided by  $N$ , we will have the proportions of yeses,  $p_1$  and  $p_2$ , respectively, for the first (or pre-) and the second (or post-) set of responses. (Note: the right-hand part of Table 6 is obtained by dividing the 8 frequencies in the left-hand part by  $N$ .)

Before proceeding to develop a scheme for testing the statistical significance of the difference between the proportions,  $p_1$  and  $p_2$ , let us note that  $p_1$  and  $p_2$  can differ only in case the frequency  $A$  differs from the frequency  $D$ , since  $p_1 = (A + B)/N$  and  $p_2 = (B + D)/N$  have  $B$  in common. Our null hypothesis is that the movie produces no change, i.e., that if the movie could be shown to the entire defined population, the proportion of yeses before and after would be exactly the same. This does not mean that an individual can't change, but it does mean that the number of changes from yes to no balances off the number of changes from no to yes. Thus we come to the proposition that on the basis of the null hypothesis we would expect those individuals who gave a changed response to split 50-50 as to direction of change. Stated differently, we would expect 1/2 of the  $A + D$  individuals (the changers) to change from yes to no and 1/2 of them to change from no to yes.

Since this is precisely analogous to tossing  $A + D$  coins, we would expect successive samples to yield a chance distribution for the number of no to yes changes which would follow the binomial with  $M = (A + D)/2$  and  $\sigma = \sqrt{(A + D)(.5)(.5)}$ ; that is, with  $n = A + D$  and  $p = 1/2$ . Note that, for  $A + D$  fixed, the number of yes to no changes is complementary to the number of no to yes changes just as when coins are tossed the number of tails is complementary to the number of heads—we need not count both. Thus a test of the significance of the deviation of either  $D$  or  $A$  from  $(A + D)/2$  tells us whether  $D$  differs significantly from  $A$ .

For  $A + D$  small, say 10 or less, we may use the actual binomial expansion to evaluate the change, but for  $A + D$  large we will need to resort to the normal curve approximation. The latter is readily accomplished by expressing  $D$  as a deviation from  $(A + D)/2$  and dividing by  $\sigma$ , or by  $\sqrt{(A + D)(.5)(.5)}$ , which gives a critical ratio,

$$CR = \frac{x}{\sigma} = \frac{D - (A + D)/2}{\sqrt{(A + D)(.5)(.5)}} = \frac{.5D - .5A}{.5\sqrt{A + D}} = \frac{D - A}{\sqrt{A + D}} \quad (19a)$$

as a value with which to enter Table A to find the probability of as large a deviation as that obtained. If this  $x/\sigma$ , or  $CR$ , is 2.58 (or larger) the  $P = .01$  level of significance will have been reached (we are here dealing with the probability of as large a deviation irrespective of direction). When  $A + D$  is not large, say 11 to 20, our approximation will be appreciably improved by deducting .5 from the deviation of  $D$  from  $(A + D)/2$ ; this is the correction for continuity again.

If we wished to carry our computations through on the basis of proportions we could express  $D$  as a proportion of  $A + D$  (similar to what we did on p. 53) but, as we shall see, it is more appropriate to introduce the sample size,  $N$ , into the picture. Dividing both numerator and denominator of (19a) by  $N$  will not change the value of the fraction, that is,

$$CR = \frac{x}{\sigma} = \frac{D/N - A/N}{\sqrt{\frac{A + D}{N^2}}}$$

If we let  $a = A/N$  and  $d = D/N$ , this may be written as



$$CR = \frac{x}{\sigma} = \frac{d - a}{\sqrt{\frac{a + d}{N}}} \quad (19b)$$

This form for  $x/\sigma$  will make more sense if we again consider Table 6, particularly the right-hand part. Note that since  $a + b = p_1$  and  $b + d = p_2$ , it follows that  $d - a = p_2 - p_1$  and accordingly a test of the significance of  $D$  as a deviation from  $(A + D)/2$  is also a test of the significance of the difference between the proportions of yeses obtained on the 2 occasions.

The denominator of the right-hand side of (19b) must be a standard deviation. Of what? Actually it is the standard deviation of the theoretical sampling distribution of differences between proportions, each difference being based on 1 sample of size  $N$ . Such a standard deviation, as we have noted previously, is referred to as a standard error. Thus we have

$$\sigma_{D_{p(r)}} = \sqrt{\frac{a + d}{N}} \quad (20)$$

as the standard error of the difference between correlated proportions. The subscript  $r$  has been added to indicate that this formula holds for related or correlated proportions. The relationship, or correlation, concept needs a brief word of explanation. If, by chance sampling,  $p_1$  were lower than the population value we would expect  $p_2$  also to be somewhat low; if  $p_1$  were by chance high we would expect  $p_2$  to be somewhat high; if  $p_1$  were near the population value (near average) we would expect  $p_2$  to be near average. This varying together is referred to as a co-relationship or correlation. Stated differently, we would not expect the 2 proportions to vary independently of each other for successive samples.

The proportions need not be based on the same individuals to be correlated. For example, if we were interested in sex differences in opinion we might randomly choose families and then ascertain the proportion of yeses among the husbands and also among the wives; for successive samplings the 2 proportions might be correlated because of a possible tendency for husbands and wives to agree on the given issue. As a second example, consider the setup involving the pairing of individuals for the purpose of having comparable experimental and control groups. The fact of pairing signifies that



the 2 groups have not been drawn independently in the sampling sense; hence there might be a tendency for the proportions based on the 2 groups to be more or less alike. (About pairing we will have more to say in the next chapter.)

Another instance for which formulas (19a), (19b), and (20) are applicable is the problem of judging the significance of the difference between proportions of yeses for 2 different questions asked of the same sample of  $N$  cases. Since the responses to the 2 questions might tend to vary together there could be a correlation between the proportions on successive samplings.

In each of the foregoing situations we have pairs of responses, and our tabulation must follow the scheme set forth in Table 6; i.e., our tabulation will lead to the frequency of yes-no, yes-yes, no-no, and no-yes responses.

Formulas (19a), (19b), and (20) are usable in other situations. When judging whether or not 2 test items differ significantly in difficulty we ordinarily have pass-fail data for both items on the same sample of  $N$  cases. Our tabulation leads to the frequencies for pass-fail, pass-pass, fail-fail, and fail-pass. The kind of response is irrelevant—it need only be such that a dichotomy is involved for each item or question.

These 3 formulas may be safely used for any size sample provided  $A + D$  is 10 or more. If  $A + D$  is less than 10, the binomial expansion provides an easily computed test of significance leading to an exact probability for as great a difference between the proportions as that observed. The  $P$  so obtained needs to be doubled to get the probability for as great a difference irrespective of direction; otherwise it is the probability for as large a difference in one direction only. About this we shall have more to say later under the heading, "One-tailed vs. two-tailed tests," p. 62.

**Independent proportions.** It is not easy to build up a formula for evaluating the difference between 2 proportions based on 2 independent samples. It can be shown mathematically that the standard error of the difference between proportions based on independent samples is given by

$$\sigma_{D_{p(1)}} = \sqrt{\frac{p_c q_c}{N_1} + \frac{p_c q_c}{N_2}} = \sqrt{p_c q_c \left( \frac{1}{N_1} + \frac{1}{N_2} \right)} \quad (21)$$

in which  $p_c$  and  $q_c$  are the proportions in the 2 categories for the 2

groups combined. The value of  $p_c$  is readily obtained by combining the 2 frequencies of yeses (or whatever the given category is) and dividing by  $N_c = N_1 + N_2$ , and as usual  $q_c = 1 - p_c$ . An observed difference divided by  $\sigma_{D_{p(i)}}$  will give an  $x/\sigma$ , or  $CR$ , interpretable as a unit normal curve deviate provided the  $N$ 's are not too small and  $p_c$  is not too extreme. The rule-of-thumb is that  $p_c$  or  $q_c$  (whichever is smaller) times  $N_1$  or  $N_2$  (whichever is smaller) shall exceed 5. When this product is between 5 and 10, a correction for continuity should be incorporated. This may be done by reducing the numerical (absolute) value of the difference,  $p_1 - p_2$ , by the quantity  $\frac{1}{2} \left( \frac{1}{N_1} + \frac{1}{N_2} \right)$ .

### SOME GENERAL CONSIDERATIONS

Before going further we should stop long enough to delineate the general problem of hypothesis testing, discuss the question of one-tailed vs. two-tailed tests, and consider the problem of what level of significance to adopt.

**Which hypothesis?** In general, successive samplings will yield a sampling distribution of frequencies or of proportions or of differences between statistical measures or certain ratios (such as  $x/\sigma$  or  $CR$  or other ratios, to be discussed later). Hypotheses, whether statistical or research, are usually concerned either with differences or with deviations. By research hypothesis we mean the hypothesis set up on the basis of theory or prior observation or on logical grounds. Such a hypothesis usually involves a prediction regarding the outcome of an experiment. By statistical hypothesis we usually mean a null hypothesis set up for the purpose of evaluating the research hypothesis.

When we are considering possible differences the null hypothesis, frequently symbolized as  $H_0$ , is pitted against an alternate hypothesis,  $H_1$ . Now  $H_0$  specifies that, for example,  $p_{pop(1)} = p_{pop(2)}$  or that 2 population values do not differ, whereas  $H_1$  might specify that  $p_{pop(1)} > p_{pop(2)}$  or that  $p_{pop(1)} < p_{pop(2)}$  or that  $p_{pop(1)} \neq p_{pop(2)}$ . Which of these alternates is appropriate depends on the research hypothesis to be tested by experiment or what question is to be answered by experiment. An experiment is carried out which yields sample values,  $p_1$  and  $p_2$ , and the difference between  $p_1$  and  $p_2$  is used to test  $H_0$  against  $H_1$ ; that is, on the basis of the

obtained difference we are to make a decision as to whether  $H_0$  or  $H_1$  is true.

Now if  $H_0$  is true we can specify the probability of obtaining by chance a difference as great as  $p_1 = p_2$  or as great as  $p_2 = p_1$  or as great as the numerical (irrespective of sign) difference,  $p_1 = p_2$ . Let  $\alpha$  represent a chosen level of significance—any level such as  $P = .10$  or  $P = .05$  or  $P = .01$  or  $P = .001$ . We reject  $H_0$ , the null hypothesis, if the probability of the obtained result is as small as the chosen  $\alpha$ , and this rejection implies the acceptance of  $H_1$ . If  $\alpha$  is not reached we accept  $H_0$ , but this acceptance merely says that  $H_0$  could be true—any of a whole series of differences near zero could also be true. This acceptance-rejection business involves risks, to be discussed below under "Choice of level of significance."

**One-tailed vs. two-tailed tests.** The 3 possible alternates listed above for  $H_1$  have to do with hypotheses admissible on the basis of either the research hypothesis or the question for which we seek an answer by way of an experiment. In general if  $H_1$  states that  $p_{1-2}$  does not equal  $p_{1+2}$  a two-tailed test is in order; if  $H_1$  specifies which population value is the larger, a one-tailed test is used. The issue as to whether we should use a one-tailed test or a two-tailed test depends on whether the scientific hypothesis being tested or at times the practical decision to be made demands that we be concerned with chance deviations in just one direction or in both directions. For situations in which we wonder whether a performance is better than chance, as in blindfold cigarette discrimination, we are concerned only with results in one direction, since any performance in which the subject is successful on less than 50 of the trials leads us, without further statistical ado, to accept the hypothesis that he can't discriminate better than chance. Thus a one-tailed test is appropriate. But for situations in which we seek to decide whether a population is split 50-50 on some question, we need to consider chance-sampling deviations in both directions; hence we should use a two-tailed test.

Next consider the problem of testing the significance of the difference between 2 proportions. If, for example, we have the proportion of votes for some question for a sample of Republicans and for a sample of Democrats as a basis for deciding whether Republicans and Democrats differ on the given issue, we would need to use a two-tailed test—we reject the hypothesis of no differ-

ence in case the obtained difference, irrespective of direction, has a probability of occurrence which is as small as  $\alpha$ , the chosen level of significance. A one-tailed test would be utilized for judging significance in an experiment in which, for example, we were trying a new drug to see if it is better as a preventative than some commonly used drug. The decision to adopt the new drug is made only in case the new drug leads to a greater proportion of immunities. Results in only one direction are crucial to the decision to change drugs. But if we were trying out 2 drugs with the idea of adopting the one which is most promising we would use a two-tailed test since significance in either direction is the basis for decision.

It is sometimes argued that whenever one predicts on the basis of theory or previous observation the outcome of an experiment, a one-tailed test is appropriate since some benefit should accrue to the researcher who has predicted the direction of the results as opposed to the investigator who, though obtaining similar results, has not predicted the outcome. The benefit comes about in that the  $t$  or  $F$  or  $\chi^2$  for, say, the  $P = .01$  level of significance need reach only 2.33 for a one-tailed as compared with 2.58 for a two-tailed test. For the  $P = .05$  level the respective values are 1.64 and 1.96. In other words a difference, to be significant, does not have to be as large for a one-tailed as for a two-tailed test. Since the situation involving prediction is equivalent to taking  $H_1$  as the hypothesis that the difference between 2 population values is in a specified direction, it is not only defensible to use a one-tailed test but actually better in the sense that if there is a real difference in the predicted direction it will be more apt to be detected by a one-tailed than by a two-tailed test. However, a few words of caution are in order.

First, the prediction should be made prior to the collection of data, that is, independently of the data to be used in testing the hypothesis. Second, one must be on guard against bad instances can be cited where an investigator after making a series of one-tailed tests failed to shift to a two-tailed test when he should have. Third, in case the results are significant in the direction opposite to the prediction one must, in effect, have a red face because the outcome is not consistent with either of the admissible hypotheses (no difference as set forth by the null hypothesis or a difference in the predicted direction as set forth by the research hypothesis being tested). It is one thing to have results which

simply fail to support a hypothesis, and quite a different thing to have an outcome which is diametrically opposed to the hypothesis.

**Choice of level of significance.** How large should  $x/\sigma$ , or  $CR$ , be before one claims significance? Asked differently, How does one choose  $\alpha$ , the value of  $P$  to be required for judging significance? There is no one answer to this question. For a long time psychologists insisted on a  $CR$  of 3.00 (equivalent to  $P = .003$  level for two-tailed test) as a rule-of-thumb value for judging significance. There might be occasions when one would desire the assurance represented by a  $P$  of .003, but it should be noted that the acceptance of the null hypothesis whenever  $CR$  does not reach 3.00 may lead too frequently to another type of erroneous conclusion. To understand this, we must consider what it means when an observed difference does not lead to the rejection of the null hypothesis. Acceptance of the null hypothesis does not prove that no difference exists. For example, a difference of 1 per cent, in number of yeses for 2 samples, which yields an  $x/\sigma$ , or  $CR$ , of .8 does not prove that there is no difference in the 2 universe values—it merely indicates that the real difference could easily be zero. However, the obtained difference of 1 could be a chance departure from a real difference of .5 or 1.2 or 1.8 or any of a whole series of values near 1. In other words, the null hypothesis is one which can be rejected but can never be proved; therefore to accept it too often because we insist on a high level of significance for rejection means that we are too apt to overlook real differences. This, plus the fact that we don't ordinarily need the assurance represented by a significance level of .003, would suggest that a  $CR$  of 3.00 is too high.

At the other extreme, a few are willing to accept as significant a difference which is 1.5 times its standard error. Since  $P = .13$  (two-tailed) for a  $CR$  of 1.5, it is readily seen that such persons would all too frequently have their publics believing that chance differences are real. A less lax level, which has had general acceptance by psychologists recently, is represented by a  $P$  of .05. This also may be a rather low level of significance for announcing something as "fact." Those writers who advocate the .05 level for research workers in psychology cite R. A. Fisher, an eminent statistician, as their authority, but they fail to point out that Fisher's applications are to experimental situations in agriculture and biology where there is far better control of sampling than is ordinarily the case in psychology.



If the findings of a study are to be used as the basis either for theory and further hypotheses or for social action, it does not seem unreasonable to require a higher level of significance than the .05 level. The answer as to what level, in terms of probability, should be adopted in order to call a finding statistically significant is not uninvolved. There is the balancing of risks: that of accepting the null hypothesis when to do so may mean the overlooking of a real difference against that of rejecting the null hypothesis when doing so may lead to the acceptance of a chance difference as real. There is the question of the likelihood of independent verification, and, finally, there is the whim of personal preference: some individuals are more eager than others to announce a "significant" finding; others are more cautious in this regard. It follows that no hard and fast rule can be given; one can interpret a given finding in terms of the probability of its occurrence by chance and then note whether the  $P$  is near the significance level adopted *prior* to the experiment because it seemed appropriate when all factors were weighed.

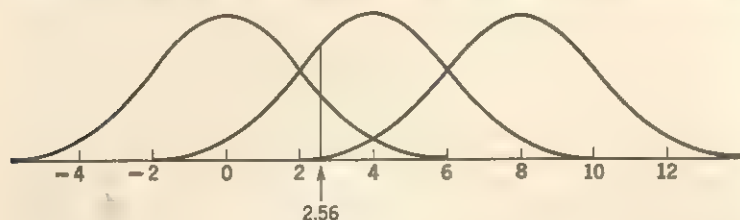
The reader will have noted from the foregoing that the testing of hypotheses involves the possibility of 2 types of erroneous conclusions. These are usually referred to as type I and type II errors, which we shall now more specifically define. Consider again the null hypothesis that no difference exists between 2 population values. If we reject this hypothesis when in fact it is true, we will have committed a type I error. If we accept the hypothesis when in fact it is false, we will have made a type II error.

The factors in choosing a level of significance might be further clarified by a somewhat different approach. Notice that when we adopt  $P = \alpha$  as our level of significance we are definitely specifying the probability of committing a type I error; it is simply  $\alpha$ . By taking  $\alpha$  smaller and smaller we can reduce the risk of making a type I error. But what happens to the probability of making a type II error as we thus reduce the risk of a type I error? The answer, and the reasoning back of the answer, can readily be understood provided one is willing to follow carefully the following line of argument. Suppose we have the proportions of immunity in 2 samples to which 2 drugs have been administered, and our question is whether drug A is superior to drug B (a one-tailed test situation). Suppose further that the standard error of the difference between the 2 proportions is .02. The exposition will be somewhat simplified if we change to percentage units—

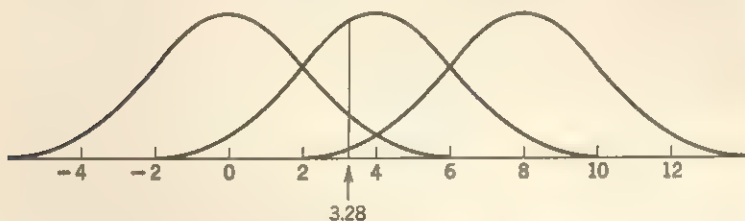


this is readily accomplished by shifting decimals for the proportions and also for the standard error; the latter becomes 2 in percentage units.

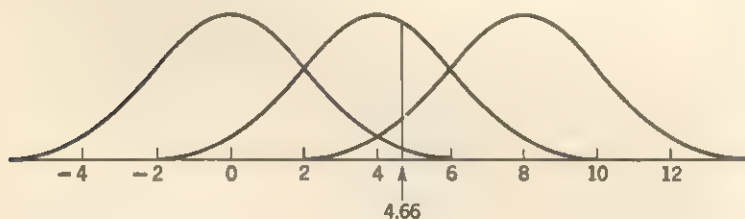
In Fig. 9 will be found a series of sampling distribution curves,



(a)  $\alpha = .10$ ,  $CR = 1.28$ .  $D$  must = 2.56.



(b)  $\alpha = .05$ ,  $CR = 1.64$ .  $D$  must = 3.28.



(c)  $\alpha = .01$ ,  $CR = 2.33$ .  $D$  must = 4.66.

Fig. 9. Type I and type II errors.

all with  $\sigma = 2$ , but with locations differing according to supposed true, or population, differences of 0, 4, and 8. The top part (a) is for  $\alpha = .10$ , the middle (b) for  $\alpha = .05$ , and the bottom (c) for  $\alpha = .01$ . In each part an ordinate has been erected at the difference required for significance at the given  $\alpha$  level of significance. These required differences spring from the fact that for a one-tailed test the  $x/\sigma$  values that cut off .10, .05, and .01 of a normal curve are 1.28, 1.64, and 2.33 respectively, and since  $\sigma$  is 2, the

respective required differences in percentages would be 2.56, 3.28, and 4.66. Sample differences falling beyond these values would be in what are termed *critical regions* for rejecting the null hypothesis at the 3 respective  $\alpha$  values. For example, values beyond 4.66 would be in the critical region when the  $P = .01$  level of significance is adopted.

From these several sampling distribution curves and with the help of a table of the normal curve functions we can specify the probability of committing a type II error for a specified (supposed) true difference. If we keep in mind that the probability of a type I error is  $\alpha$  ( $= .10, .05$ , or  $.01$ ), and that we can make a type I error only when the true difference is zero, we see that the proportionate areas beyond 2.56, 3.28, and 4.66 for the 3 curves centering at zero represent the probabilities of making a type I error for the respective  $\alpha$  values. For all sample values in the regions to the left of 2.56, 3.28, and 4.66 we would correctly accept the null hypothesis when in reality it is true. The probabilities for correct acceptance are given by  $1 - \alpha$ , or .90, .95, and .99 respectively.

Let us now consider the supposition that the true difference is 2. If 2 is the true difference, any obtained difference falling in the region to the right of 2.56, 3.28, and 4.66 will, for the respective levels of significance, lead to the *correct* decision that a true difference exists. The probabilities for these correct inferences are obtained by expressing 2.56, 3.28, and 4.66 as deviations from 2 (the supposed true value being considered), taking the deviation relative to the standard error of the difference ( $= 2$ ), and thus obtaining  $x/\sigma$  values of  $(2.56 - 2)/2 = .28$ ,  $(3.28 - 2)/2 = .64$ , and  $(4.66 - 2)/2 = 1.33$ . Looking these values up in a table of the normal curve we get probabilities, for correctly rejecting the null hypothesis, of .39, .26, and .09, for the respective specified levels of significance, *when* the true difference is 2 percentage points. Probabilities for *correctly* rejecting the null hypothesis are usually symbolized by  $\beta$ . Note that all sample values falling in the region to the left of 2.56, 3.28, and 4.66 (for the curves centering at 2) will lead to the false acceptance of the null hypothesis. The probabilities of making type II errors will correspond to the proportionate areas, for the curves centering at 2, to the left of these 3 points (when we have the one-tailed test as considered here). These probabilities will, of course, be given to us by  $1 - \beta$ .

Thus we have  $1 - .39 = .61$ ,  $1 - .26 = .74$ , and  $1 - .09 = .91$  as the probabilities of making a type II error, when the true difference is 2 and for the .10, .05, and .01 levels of significance. Note that taking  $\alpha$  smaller and smaller increases the probability of making a type II error.

For a true difference of 4 we can by a similar line of reasoning obtain the probability of correctly rejecting the null hypothesis and the probability of falsely accepting the null hypothesis, when using any one of the specified values of  $\alpha$ . These probabilities will involve the areas, under the curves centering at 4, to the right of 2.56, 3.28, and 4.66 (for the  $\beta$ 's) and to the left of these same points (for the type II errors). The student can, as an exercise, readily verify that the areas to the right of 2.56, 3.28, and 4.66 are approximately .76, .64, and .37 respectively. Subtracting each of these from unity will yield the probabilities, .24, .36, and .63, of falsely accepting the null hypothesis or committing a type II error when the true difference is 4 and for  $\alpha$ 's of .10, .05, and .01. Again, the smaller we take  $\alpha$  the larger the probability of making a type II error.

The probabilities given in the last 2 paragraphs, along with similar figures for other supposed true differences, have been assembled in Table 7. A careful study of this table reveals the general rule that the smaller the value of  $\alpha$  the smaller the probability ( $\beta$ ) of correctly rejecting the null hypothesis and the larger the probability ( $1 - \beta$ ) of committing a type II error. Thus when we reduce the probability of making a type I error by choosing  $\alpha$  small, we do so at the risk of more often making a type II error. Note also that regardless of  $\alpha$ , the probability of making a type II error decreases as the true differences deviate farther and farther from zero. This is another way of saying that the larger the true difference the more apt we are to detect it by experiment, and conversely the smaller the difference the less likely we are to discover it.

Incidentally, the value of  $\beta$  for various possible true differences is referred to as the *power* of the statistical test for detecting the difference. If we plotted the  $\beta$ 's in, say, the  $\alpha = .05$  column of Table 7 against the scale of possible differences we would have an ascending curve which would represent the *power function* of the test. It is beyond the scope of this text to consider in detail the concepts having to do with the power of a test. It should be

Table 7. PROBABILITY ( $\beta$ ) OF CORRECTLY REJECTING THE NULL HYPOTHESIS AND PROBABILITY ( $1 - \beta$ ) OF TYPE II ERROR ASSOCIATED WITH 3 LEVELS OF SIGNIFICANCE ( $\alpha$ 's OF .10, .05, .01) WHEN CERTAIN TRUE DIFFERENCES ARE SUPPOSED TO EXIST

$\alpha \rightarrow$	$\beta$			$1 - \beta$		
	.10	.05	.01	.10	.05	.01
True diff.						
1	.22	.13	.03	.78	.87	.97
2	.39	.26	.09	.61	.74	.91
3	.59	.44	.20	.41	.56	.80
4	.76	.64	.37	.24	.36	.63
5	.89	.79	.57	.11	.21	.43
6	.96	.91	.75	.04	.09	.25
7	.99	.97	.88	.01	.03	.12
8	.997	.99	.95	.003	.01	.05
9	> .999	.997	.975	< .001	.003	.025
10	> .999	> .999	.996	< .001	< .001	.004

remarked, however, that statistical tests differ in their power, and to understand this we would need to have more information regarding various tests that might be used to test a given research hypothesis. For instance, power depends upon the choice of the critical region for rejecting the null hypothesis—for the first drug problem considered above, a one-tailed test is more powerful than a two-tailed test. In the next chapter we will be considering, among other things, differences between averages or central values, at which time it will be found that a test based on comparing means will be more powerful than one based on medians.

Perhaps the discerning student will have noted that increasing sample size (or sizes) tends to reduce standard errors. In the above discussion we supposed that we had  $N$ 's and proportions such that the standard error of the difference ( $\sigma_D$ ) was 2 percentage units. Quadrupling the  $N$ 's would reduce the  $\sigma_D$  to 1 percentage unit. How would this affect the results deduced from Fig. 9 and set forth in Table 7? Take, for example,  $\alpha = .01$  and suppose a true difference of 2 percentage points. With  $\sigma_D = 1$ , an obtained

difference would have to fall in the region beyond  $2.33 \times 1 = 2.33$  to be judged significant at the .01 level. With a true difference of 2, the proportion of sample values falling beyond 2.33, calculated by taking  $(2.33 - 2)/1 = .33 = x/\sigma$ , is found to be .37. This is a  $\beta$  value to be contrasted with a  $\beta$  of .09 given in Table 7. We see, therefore, that quadrupling the sample  $N$ 's has increased fourfold the probability of detecting a difference of 2 points. Or stated differently, the probability of a type II error has been reduced from .91 to .63. The moral is plain: one way of reducing the risk of making a type II error, without increasing the risk of a type I error, is to increase  $N$  or  $N$ 's. Whether this is feasible will usually depend upon the resources available to the investigator.

Although contemporary mathematical statisticians usually consider hypothesis testing in terms of a definite reject-accept decision according to whether the chosen level of significance is or is not reached, there is another possibility. One might follow the rule of rejecting the null hypothesis when  $P$  is less than .01 (say), accepting it when  $P$  is greater than .10, and reserving judgment when  $P$  is between .10 and .01. This, in effect, introduces a region of indecision, or calls for a postponement of decision until the experiment is repeated or more data are collected. Another possibility, when a decision is not required for some practical reason, is simply to report that a difference is significant at the .09 or the .04 or the .002 or whatever level is reached, and then let the reader evaluate the finding according to his own preferred level of significance (which he is apt to do anyway unless he is too naive).

There are a couple of other points regarding significance. First, a statistically significant difference doesn't necessarily mean a difference either of practical significance or of scientific import. Sometimes a "what of it" is not an impertinence. Second, the habit of merely checking to see whether a result reaches a chosen level of significance should not lead one to overlook the possibility of claiming, when appropriate, that a much higher level of significance was attained than the preresearch chosen level.

### SUMMARY

In this chapter we have given a brief account of the concept of probability and have sketched procedures for applying proba-

bility notions in the testing of hypotheses involving frequencies and proportions (or percentages). We have noted the conditions for which it is safe to use an  $x/\sigma$  (or  $CR$ ) and the normal curve to approximate probabilities. If these conditions do not hold (when samples are small or proportions are extreme), one can obtain  $P$  exactly by way of the actual binomial expansion for situations involving one proportion and for two correlated proportions. For proportions based on independent samples exact  $P$ 's may be ascertained by another, and more complicated, method to be presented later (p. 240).

The discussion of this chapter is only an introduction to the theory of statistical inference, or the use of probability in the testing of hypotheses. We have, however, developed the general principles. The extension of the theory to hypotheses involving continuous variables for relatively simple situations will be given in the next 2 chapters, with methods for more complex situations being postponed to later chapters (14, 15, 16, 17, 18). In Chapter 13 we shall discuss more extensive procedures for handling hypotheses regarding frequencies and proportions.



## CHAPTER 6

### Inference: Continuous Variables

As will be recalled, a frequency distribution for measurements on a continuous variable is describable with respect to central value, variability, skewness, and kurtosis; hence hypotheses involving continuous variables will be concerned with at least 1 of the descriptive measures of these 4 features of a frequency distribution. To test a given hypothesis we need information regarding the sampling behavior of the descriptive measure being used (or of some ratio containing the measure).

In the previous chapter we were able to make certain easy deductions. We saw, at the intuitive level, that the sampling distribution of proportions and of differences between proportions tends, under specified conditions, to be normal in distribution with specifiable standard deviation, and that we could set up a deviation,  $x$ , such that the ratio  $x/\sigma$  tends to follow the unit normal curve. Unfortunately, the behavior of random sampling distributions of the measures that describe frequency distributions cannot so readily be determined. Accordingly, we will need to lean on the deductions made by the mathematical statistician, who has the task of ascertaining mathematically the characteristics of random sampling distributions when certain conditions and assumptions hold. We can learn how to use his results as a basis for testing hypotheses without necessarily understanding his mathematical derivations.

Since hypotheses involving means arise frequently in practice and since inferences based on means serve to illustrate further the general theory of statistical inference, we shall give considerable detail on sampling errors connected with means. We shall present first an easily duplicated demonstration of the chance

variation of means, and then a discussion of some theory and its use as a basis for hypothesis testing. This chapter will be restricted to the large sample situation, with requisite sample size specified at appropriate times.

### EMPIRICAL DEMONSTRATION

The operation of chance sampling errors for means and standard deviations can be illustrated by tossing, say, 7 coins 50 times and tabulating the number of heads per toss. The obtained frequencies will usually vary somewhat from those expected, which would be proportional to 1, 7, 21, 35, 35, 21, 7, 1 (as obtained by the binomial expansion). When the mean number of heads for 50 tosses is computed, it is not likely to be exactly 3.5 ( $np$ , the mean of the expected distribution), and the discrepancy from 3.5 can be attributed to chance. Likewise, 100 tosses will show departures from the expected frequencies, and consequently the mean based on 100 tosses will differ more or less from 3.5. Furthermore, and for the same reason, the standard deviation of the obtained distribution of heads will likely differ from 1.323 ( $\sqrt{npq}$ , the  $SD$  of the expected frequencies). The student, as an exercise, can demonstrate the foregoing statements by actually tossing coins. Indeed it will be quite instructive if each class member tosses 7 coins 50 times, each time tallying the number of heads that turn up. This will lead to a frequency distribution running (possibly) from 0 to 7 heads, with an  $N$  of 50. Then a second series of 50 tosses should be made, thus providing a second distribution. The 2 frequency distributions can be combined, so each student will have 3 distributions, 2 with  $N$ 's of 50 and 1 with an  $N$  of 100. Note that chance is so operating as to produce a distribution somewhat similar to the expected, but at the same time is operating in such a manner as to lead to discrepancies between observed and expected frequencies.

Each student should compute the means and the standard deviations for each of the 3 distributions. Note how far these values depart from the expected mean of 3.5 and the expected standard deviation of 1.323. Then the several means and standard deviations secured by the class members should be brought together. In order better to understand what happens when each of several persons tosses 7 coins 50 times, i.e., takes a sample of 50 tosses, a

frequency distribution of the means, also of the  $SD$ 's, based on 50 tosses should be made. Likewise a separate distribution should be made for the  $M$ 's based on 100 tosses; also, the  $SD$ 's. A study of these distributions should provide answers to such questions as: Their central tendencies are near what values? What is the extent of dispersion for these distributions of  $M$ 's and  $\sigma$ 's? Is there any difference in the dispersion for the distribution of means based on 50 tosses and that based on 100 tosses? How would you account for this difference? In general, what is the shape of these distributions of  $M$ 's and  $\sigma$ 's?

In Table 8 will be found the distributions of the means obtained

Table 8. DISTRIBUTIONS OF 600 MEANS BASED ON 50 TOSSES AND 300 MEANS BASED ON 100 TOSSES OF 7 COINS

	50 Tosses	100 Tosses
4.00-4.09	3	
3.90-3.99	14	
3.80-3.89	35	4
3.70-3.79	50	23
3.60-3.69	98	58
3.50-3.59	119	78
3.40-3.49	120	85
3.30-3.39	85	32
3.20-3.29	52	17
3.10-3.19	21	3
3.00-3.09	2	
2.90-2.99	1	
Number of means	600	300
Mean of means	3.516	3.513
$SD$ * distribution		
of means	.190	.135
Expected $SD$	.187	.132

\* Corrected for grouping.

by several of the author's classes. Though these are not models for number of intervals, they are nevertheless sufficient as a basis for answering the foregoing questions. Note that both distributions appear to be normal, that both center very near the mean of the theoretical distribution (3.5), and that the variability for means based on 100 tosses is less than that based on 50 tosses. It would thus seem that means based on 100 tosses are somewhat more stable or less variable than those based on 50 tosses. Does

this suggest that a larger number of tosses, i.e., a larger sample, would tend to iron out the chance factors that operate to produce discrepancies between the observed distribution of number of heads and the expected distribution calculated by the binomial expansion? Do you think that means based on 500 tosses would show less dispersion than means based on 100 tosses?

According to the mathematical statisticians, the standard deviation of the distribution of means is expected to be equal to 1.323 (expected  $\sigma$  of the distribution of number of heads) divided by the square root of the sample size. Note at the bottom of Table 8 that the *SD*'s of the distributions of means, .190 and .135, are very near the expected values of .187 and .132 obtained from  $1.323/\sqrt{50}$  and  $1.323/\sqrt{100}$ , respectively.

Summarizing the results of the above empirical work, we see that the means for successive samples tend to distribute themselves normally about the expected or universe mean with a spread or standard deviation which is very near the value predicted by mathematical theory. The student should keep these empirical distributions and deductions therefrom in mind as we now proceed to a more detailed consideration of what the mathematical statistician says will happen when successive samples of a given size are drawn from a defined universe or population or supply.

### MORE SAMPLING THEORY

The discussion here holds for what is known as simple *random sampling*. As specified in the previous chapter, the conditions for simple random sampling are that the sample should be drawn in such a way that each individual (person, plant, animal, etc.) in the defined universe shall have an equal chance of being included in the sample, and that the drawing of one individual shall in no way affect the drawing of another. The aim is, of course, to obtain a sample which will, within limits of random or chance errors, be representative of the universe from which it was drawn.

Let

$N$  = the number of cases, or size of sample.

$M$  = the mean of any sample (known, i.e., computed).

$\sigma$  = the *SD* of any sample (known, i.e., computed).

$M_{pop}$  = the mean of the defined population (unknown).

$\sigma_{pop}$  = the *SD* of the defined population (unknown).

The  $M_{pop}$  and  $\sigma_{pop}$  are for the distribution of scores or measurements for all the individuals in the defined universe. It is not assumed that this universe distribution is exactly normal; it may be skewed slightly. Strictly speaking, the number,  $N_{pop}$ , of cases in the universe should be infinitely large, but failure to meet this requirement is not serious. As will be seen later, the adjustment necessary when a sample of  $N$  cases is drawn from a limited (finite) universe of  $N_{pop}$  cases is of the order of  $N/N_{pop}$ ; if it is known that  $N_{pop}$  is very large relative to  $N$ , the formulations about to be presented will be sufficiently accurate for all practical purposes.

Now suppose we draw a sample of  $N$  cases, compute the mean and standard deviation, then draw another sample of the same size and compute its mean and standard deviation, and so on until a large number of samples, say 10,000, have been drawn. We will then have 10,000 means and 10,000 standard deviations, each based on  $N$  cases. When we make a distribution of the 10,000 means and of the 10,000 standard deviations, we have random sampling distributions. From the viewpoint of mathematical rigor, the number of successive samples should be much larger than 10,000, certainly far larger than the 600, or 300, successive samples of Table 8, in which we have only the beginning of 2 random sampling distributions.

By rather complex mathematical methods it can be shown that, if successive samples of constant size,  $N$ , are drawn randomly from a normally distributed universe or population with mean equal to  $M_{pop}$  and  $SD$  equal to  $\sigma_{pop}$ , the successive sample means will be normally distributed about  $M_{pop}$ , and the standard deviation of this sampling distribution will be  $\sigma_{pop}/\sqrt{N}$ . The random sampling distribution of the successive standard deviations will center at  $\sigma_{pop}$  (there is a small bias here which need not concern us at this time). For  $N$  large (100 or more) this distribution of  $\sigma$ 's will be approximately normal with standard deviation equal to  $\sigma_{pop}/\sqrt{2N}$ . These mathematical findings have often been checked empirically. Our Table 8 provides a limited check on the sampling theory regarding the mean.

We are now in position to consider a term used in the previous chapter. In general, the *standard error* of a statistical measure is the standard deviation of the sampling distribution for the given measure. The square of the standard error is called the *sampling variance*. For the practical statistician, the sampling distribution

is hypothetical, and hence its standard deviation must be determined by a different formula from that used for computation from an actual distribution. The value given by  $\sigma_{pop}/\sqrt{N}$  is called the standard error of the mean and may be designated as  ${}_t\sigma_M$ , the subindex  $t$  is used to indicate "true" value (not estimated). Each sample mean can be expressed in standard form (analogous to standard scores) as  $(M - M_{pop})/{}_t\sigma_M$ , and these relative deviates will form a normal distribution with mean of zero and standard deviation of unity. By reference to Table A we can readily specify the chances of obtaining a sample mean yielding a deviation as great as that for a given  $M$ , provided the value of  $M_{pop}$  is known. But in practical work  $M_{pop}$  is the unknown about which we desire to make an inference on the basis of just 1 sample.

Before resolving this practical problem, we must call attention to the fact that the universe standard deviation,  $\sigma_{pop}$ , needed to obtain  ${}_t\sigma_M$  is also an unknown. A single sample will yield a standard deviation,  $\sigma$ , which, being a sample value, will of course deviate more or less from  $\sigma_{pop}$ . In order that an inference about  $M_{pop}$  may be made from a single sample,  ${}_t\sigma_M$  is estimated by using  $\sigma/\sqrt{N}$ ; i.e., the unknown  $\sigma_{pop}$  is replaced by the sample  $\sigma$  as an estimate. Instead of the true value for the standard error of the mean as given by  $\sigma_{pop}/\sqrt{N}$ , we have an approximate value,  $\sigma/\sqrt{N}$ . Let  $\sigma_M$ , defined as  $\sigma/\sqrt{N}$ , stand for the approximate standard error.

The ignorance concerning  $\sigma_{pop}$ , and the consequent approximate value for the standard error of a given mean, lead to a reconsideration of the sampling distribution of means expressed as relative deviates. As already pointed out, the means from successive samples will be distributed normally, and the relative deviates,  $(M - M_{pop})/{}_t\sigma_M$ , will likewise be distributed normally since  ${}_t\sigma_M = \sigma_{pop}/\sqrt{N}$  is a constant. When (as is nearly always the case) we have  $\sigma$  instead of  $\sigma_{pop}$  and wish to make an inference about a universe mean, we need to know something of the sampling behavior of successive sample means expressed as relative deviates from  $M_{pop}$  where  $\sigma_M$  is not a constant but varies from sample to sample because the several sample standard deviations vary. Thus the relative deviate of the first sample mean will be  $(M_1 - M_{pop})$  divided by  $\sigma_1/\sqrt{N}$ ; for the second sample,  $(M_2 - M_{pop})$  divided by  $\sigma_2/\sqrt{N}$ ; and so on. The distribution of these relative deviates will *not* approximate normality unless



$N$  is fairly large. Thus the use of an estimate of  $\sigma_{pop}$  in determining  $\sigma_M$  imposes the restriction that  $N$  shall not be too small. If  $N$  is not less than 30, we can safely use the normal curve as the basis for drawing an inference or testing a hypothesis regarding  $M_{pop}$ . This chapter's discussion of sampling is therefore not applicable unless  $N$  is greater than 30. The refinements necessary for  $N$ 's less than 30 will be given in the next chapter.

### HYPOTHESES REGARDING A SINGLE MEASURE

Whether the foregoing theory is used as a basis for making an inference about a population value or for testing some hypothesis depends on the practical problem faced by the investigator. We shall now consider hypothesis testing, and later we shall discuss a type of inference which is useful both when we do and do not have a research hypothesis in mind.

**Single mean.** The procedure for testing a hypothesis about a population mean on the basis of a sample mean (and  $SD$ ) for  $N$  cases is very similar to that for testing a hypothesis when we have a sample proportion (discussed earlier, pp. 52-55). We let  $M_h$  stand for a hypothesized value of  $M_{pop}$ . Our sample mean,  $M$ , taken as a deviation from  $M_h$ , is expressed in the form of an  $x/\sigma$ , that is, as  $(M - M_h)/\sigma_M$ . The theory tells us that if  $M_h$  is true (i.e., corresponds to  $M_{pop}$ ), successive sample  $M$ 's will be distributed normally about  $M_h$  with standard deviation =  $\sigma_M$  (approximately). In testing the given hypothesis we are merely raising the question as to whether it is reasonable to believe that our observed sample mean,  $M$ , belongs to a sampling distribution centering at  $M_h$ . Put differently, does  $M$  deviate significantly from  $M_h$ ? To answer this we need to know the probability of as large a deviation on the basis of chance sampling errors, and to get this probability we need only enter Table A with  $(M - M_h)/\sigma_M$  as an  $x/\sigma$  (or  $CR$ ). If we have decided to adopt the  $P = .01$  level for judging significance, we reject the hypothesis when  $(M - M_h)/\sigma_M$  reaches 2.58 (for a two-tailed test) or 2.33 (for a one-tailed test); otherwise we accept the hypothesis.

Actually, there are relatively few occasions in psychological research for which either scientific theory or prior observation provides us with a hypothesis concerning the mean for a population on some variable. An exception is the mean of changes, to be discussed shortly.

As an example of a situation for which the testing of a hypothesis about a mean is appropriate we cite the IQ tests. For reasons which we shall not discuss here, a properly constructed test should yield 100 as the average IQ for the population of children for any given age level. Consider Form L of the 1937 Revision of the Stanford-Binet Scale. For age 7 a sample of 202 gives a mean of 101.78 and a  $\sigma$  of 16.18. The value of  $\sigma_M$  becomes  $16.18/\sqrt{202} = 1.14$ . From these figures we have  $(M - M_h)/\sigma_M = (101.78 - 100)/1.14 = 1.56$  as an  $x/\sigma$ . Turning to Table A we find that the  $P$  for as large a deviation (irrespective of direction—a two-tailed test is needed here) from 100 is .12. Since this probability is not as small as our arbitrarily chosen  $P = .01$  level of significance, we accept the hypothesis that the 1937 Stanford-Binet meets the requirement of yielding an average of 100 at age 7. That the scale is not entirely satisfactory in this regard is evident when we consider the  $M$  of 104.28 and  $\sigma$  of 16.42 for a sample of 204 nine year olds. We have  $\sigma_M = 1.15$ , which leads to an  $x/\sigma$  of  $(104.28 - 100)/1.15 = 3.72$ . Since the probability of as large a deviation is about .0002, we reject the hypothesis that the scale would yield a population mean of 100 at age 9.

**Significance of mean change.** A frequently encountered problem is that of evaluating changes in order to say whether some provided experience or change in conditions leads to a shift in performance.

Let

$X_1$  = score prior to experience (or under one condition).

$X_2$  = score after the experience (or under second condition).

$D = X_2 - X_1$  = change score.

Or we might take  $D = X_1 - X_2$  if losses instead of gains are of interest, but regardless of which way we define the  $D$  score, the subtraction is made in the same direction for all  $N$  cases and negative signs are kept. A sample of  $N$  individuals will give us  $N$  changes, or  $N$   $D$ 's. We can either make or conceive of a distribution of the  $D$ 's. This distribution will have a mean,  $M_D$ , and a standard deviation,  $\sigma_D$ , whence we can get the standard error of the mean difference:  $\sigma_{M_D} = \sigma_D/\sqrt{N}$ . In other words, a mean change is treated just like any other mean. Regardless of one's hunch or prediction about the effect of the experience (or the effect of the change in conditions), one sets the null hypothesis that there is no effect. This is equivalent to saying that, if we

had  $X_1$  and  $X_2$  scores on the defined population, the value of  $M_D$  would be zero. If this hypothesis is true and if we were to take successive samples of size  $N$  we would expect that the sample means would be distributed normally about zero with  $SD = \sigma_{M_D}$ . To test the null hypothesis we simply take our obtained  $M_D$  as a deviation from the null value of zero and divide by  $\sigma_{M_D}$ . That is,  $(M_D - 0)/\sigma_{M_D} = M_D/\sigma_{M_D}$ . This  $x/\sigma$  is then used as an entry into Table A in order to specify the probability of as large a mean difference as our sample  $M_D$  arising solely on the basis of chance sampling. Whether we reject or accept the hypothesis of no effect depends on whether  $P$  does or does not reach the chosen level of significance. We should use a one-tailed test here if the research hypothesis predicted the direction of the change, but if we had no a priori hypothesis as to the direction of change we would need to use the two-tailed test.

A word should be inserted about the required computations since there is some danger of confusion when one is confronted with the calculation of  $M$  and  $\sigma$  for scores (changes) which are both positive and negative, and sometimes zero. The gross score formula for the mean (2) and that for the standard deviation (6b) are applicable provided one takes  $\Sigma D$  (equivalent to  $\Sigma X$ ) as the algebraic sum. The equivalent of  $\Sigma X^2$ , that is,  $\Sigma D^2$ , raises no problem since the squaring process automatically eliminates negative signs. There are two reasons why one should make a frequency distribution of the  $D$ 's. First, the theory assumes that the  $D$ 's approximate a normal distribution; if a distribution is made one has at least a rough check on this assumption (there are statistical methods for checking this assumption; see p. 82 and also p. 236). Second, if  $N$  is sizable, computation from a frequency distribution is more economical of time than use of the gross score formulas. In laying out the intervals, one must provide a place for tabulating zero  $D$ 's. This can conveniently be accomplished by the following illustrative scheme which includes only the 4 intervals near zero: 2-3, 0 1, -1-2, -3-4 (for  $i = 2$ ); 3 5, 0-2, -1-3, -4-6 (for  $i = 3$ ); 4 7, 0 3, -1-4, -5 8 (for  $i = 4$ ); etc. Note that the last given intervals in each set are for negative  $D$ 's.  $AO$  taken as the midpoint of the bottom interval will be a negative number, and must be treated as such when entered into formula (3).

**Other single measures.** The general theory of statistical inference may be extended to testing hypotheses concerning any

descriptive measure provided information is available (from the mathematical statistician) concerning the characteristics of the random sampling distribution of the measure. When the sampling distribution is normal in form with known or estimable variability, we may proceed to test hypotheses by setting up an  $x \sigma$  (or  $CR$ ). For this purpose we need formulas for the standard errors of different measures. The formulas about to be presented are based on the assumption that the score distribution is normal or approximately so.

As previously noted, for  $N$  greater than 30 we may safely use

$$\sigma_M = \frac{\sigma}{\sqrt{N}} \quad (22)$$

as the standard error of the mean. For  $N$  greater than 100 it is safe to take

$$\sigma_{mdn} = \frac{1.253\sigma}{\sqrt{N}}$$

as the standard error of the median. A comparison of the standard error of the mean with that of the median indicates that the mean fluctuates less than the median; i.e., the mean is a more stable measure of central value than the median. In order to reduce the standard error of the median to the same magnitude as that of the mean it is necessary to take 57 per cent more cases, i.e., increase  $N$  by 57 per cent. It follows from this that the use of the median for distributions which are reasonably normal in form is equivalent to throwing away a large proportion of the cases.

The sampling errors involved in measures of dispersion are

$$\begin{aligned} \sigma_\sigma &= \frac{\sigma}{\sqrt{2N}} = \frac{.707\sigma}{\sqrt{N}} = .707\sigma_M \\ \sigma_{AD} &= \frac{.756(AD)}{\sqrt{N}} \\ \sigma_Q &= \frac{1.166(Q)}{\sqrt{N}} \end{aligned} \quad (23)$$

From these error formulas it will be seen that, considering the error relative to the magnitude of the measures of dispersion,  $\sigma$

is the most stable measure of variation. Provided  $N$  is 100 or more, the sampling distributions for these measures of dispersion are such that their standard errors can be utilized in exactly the same way as the standard error of the mean.

The standard errors for measures of skewness and kurtosis, as defined on p. 28, are

$$\sigma_{g_1} = \sqrt{\frac{6}{N}} \quad (24a)$$

$$\sigma_{g_2} = \sqrt{\frac{24}{N}} = 2 \sqrt{\frac{6}{N}} = 2\sigma_{g_1} \quad (24b)$$

These 2 formulas are based on the assumption that the sample has been drawn from a normally distributed population, and therefore they can be legitimately used in testing the assumption of normality. It will be recalled that for normal distributions both  $g_1$  and  $g_2$  are equal to zero, but for a sample they may not be zero, however sample values should not show a greater deviation from zero than can be reasonably attributed to chance. If a sample yields a  $g_1$  value which is more than say 2.58 times its sampling error one would suspect that the sample was not drawn from a symmetrically distributed supply. Likewise if  $g_2$  deviates more than 2.58 times its standard error, one would question whether it is reasonable to believe that the population or supply is distributed with normal kurtosis. A two-tailed test is appropriate here, and consequently choosing 2.58 is equivalent to adopting the .01 level of significance.

### HYPOTHESES ABOUT DIFFERENCES

One of the foremost problems in practical statistics is the comparison of group trends. We may wonder whether one college group is superior to another, whether practice on a task improves performance, whether rats learn more rapidly when food or when water is the incentive, whether reaction time is faster to sound than to light, whether the sexes show a difference in variational tendency, whether one learning method is better than another, etc. In order to answer questions like the above it is necessary to make observations on samples from 2 groups or on the same group under 2 different experimental conditions, and then to com-

pute appropriate statistical measures for the variable upon which we wish to make the comparison.

Thus, typically, we have 2 samples of  $N_1$  and  $N_2$  cases or 2 sets of scores on just  $N$  cases under 2 different conditions, with means  $M_1$  and  $M_2$  and standard deviations  $\sigma_1$  and  $\sigma_2$ , where the subscripts refer to the 2 sets of scores. As we have learned, each mean is subject to sampling fluctuations, therefore the difference between the means will also be subject to sampling fluctuations. Even though  $M_{pop-1} = M_{pop-2}$  there may be a difference between sample means because of chance sampling errors. To test an obtained difference for significance we will need a measure of the sampling error of differences, i.e., the standard error of the difference between two means. Knowing this standard error we can set up the null hypothesis that there is no difference between the two population means and then reject or accept this hypothesis according to whether the obtained difference does or does not reach an appropriate level of significance.

Here, as in the case of the difference between proportions, we must distinguish between the situation where our 2 means are based on independent as opposed to nonindependent (correlated) scores.

**Difference between correlated means.** Let us again consider the method outlined above for testing the significance of a mean change. As implied there, the  $X_1$  and  $X_2$  scores could stand for performance for  $N$  individuals under 2 different conditions. A little simple algebra at this point will lead to some interesting results. As before, we let

$$D = X_2 - X_1$$

By definition the mean of the distribution of these  $N$  difference scores will be

$$\begin{aligned} M_D &= \frac{\Sigma D}{N} = \frac{\Sigma(X_2 - X_1)}{N} \\ &= \frac{\Sigma X_2}{N} - \frac{\Sigma X_1}{N} \end{aligned}$$

hence

$$M_D = M_2 - M_1 = M_M$$

by which we see that the mean of the difference is equal to the difference between the means. This will, of course, be true for every



sample. It follows therefore that when we test the significance of  $M_D$  as a deviation from zero we are also testing the significance of  $D_M$  as a deviation from zero. In other words, we are testing the significance of the difference between 2 means based on the same  $N$  cases.

When testing  $M_D$  we calculated  $\sigma_D$ , thence  $\sigma_{M_D}$ . Let us consider a bit further the standard deviation of the distribution of differences,  $\sigma_D$ . We first express the  $D$ 's as deviations from their own mean, i.e.,  $d = D - M_D$ . Since  $D = X_2 - X_1$  and  $M_D = M_2 - M_1$ , we have

$$d = (X_2 - X_1) - (M_2 - M_1)$$

which, when the parentheses are removed and the terms shifted, becomes

$$d = X_2 - M_2 - X_1 + M_1$$

or

$$d = (X_2 - M_2) - (X_1 - M_1)$$

Both these new parentheses terms define deviation units of the type  $x = X - M$ , so that  $d = x_2 - x_1$ . The standard deviation squared, or variance, of the difference can be expressed by substituting  $d$  for  $x$  in formula (4); thus

$$\sigma_D^2 = \frac{\sum d^2}{N}$$

If we replace  $d$  by its equivalent, we have

$$\sigma_D^2 = \frac{\sum (x_2 - x_1)^2}{N} = \frac{\sum x_2^2}{N} + \frac{\sum x_1^2}{N} - \frac{2\sum x_2 x_1}{N}$$

The first 2 of the 3 terms on the right are obviously the variances for the second and first sets of scores. The last term, involving the sum of the cross products of  $x_2$  and the  $x_1$  with which it is paired, has to do with the degree of correlation between, or similarity of, the scores that belong to the same individual. The reader is asked to take on faith, without further explanation here, the fact that the last term becomes  $2r_{12}\sigma_1\sigma_2$ , in which  $r$  is a measure of correlation. Hence we can write

$$\sigma_D^2 = \sigma_2^2 + \sigma_1^2 - 2r_{12}\sigma_1\sigma_2$$

or

$$\sigma_D = \sqrt{\sigma_2^2 + \sigma_1^2 - 2r_{12}\sigma_1\sigma_2}$$

Since the standard error of any mean is given by dividing the standard deviation by the square root of  $N$ , we secure the standard error of the mean difference by dividing  $\sigma_D$  by  $\sqrt{N}$ , i.e.,

$$\begin{aligned}\sigma_{M_D} &= \frac{\sigma_D}{\sqrt{N}} = \frac{\sqrt{\sigma_1^2 + \sigma_2^2 - 2r_{12}\sigma_1\sigma_2}}{\sqrt{N}} \\ &= \sqrt{\frac{\sigma_1^2}{N} + \frac{\sigma_2^2}{N} - \frac{2r_{12}\sigma_1\sigma_2}{N}}\end{aligned}$$

The first 2 terms under the last radical are the sampling variances of the 2 means, and since  $2r_{12}\sigma_1\sigma_2/N$  can be written as

$$2r_{12} \frac{\sigma_1\sigma_2}{\sqrt{N}\sqrt{N}}$$

we have finally that

$$\sigma_{M_D} = \frac{\sigma_D}{\sqrt{N}} = \sqrt{\sigma_{M_1}^2 + \sigma_{M_2}^2 - 2r_{12}\sigma_{M_1}\sigma_{M_2}}$$

Since each  $M_D = D_M$ , it follows that  $\sigma_{M_D} = \sigma_{D_M}$ , or that the standard error of the mean difference is equal to the standard error of the difference between the 2 means. Thus we have 2 ways for evaluating a difference between nonindependent means. We can compute  $M_D, \sigma_D$ ; thence

$$\sigma_{M_D} = \frac{\sigma_D}{\sqrt{N}} \quad (25a)$$

or we can compute  $M_1, M_2, \sigma_1, \sigma_2$ , and  $r_{12}$ , and then obtain

$$\sigma_{M_D} = \sqrt{\sigma_{M_1}^2 + \sigma_{M_2}^2 - 2r_{12}\sigma_{M_1}\sigma_{M_2}} = \sigma_{D_M} \quad (25b)$$

Formula (25b) is usually referred to as the standard error of the difference between correlated means, hence the symbol  $\sigma_{D_M}$ .

But by working with the difference between paired scores, we can obtain the standard error of the mean difference (= difference between means) without computing  $r$ . Even after we have learned how to compute  $r$ , it matters not whether we compute the standard error of the difference between means of related scores by formula (25b) or whether we compute its equivalent, the standard error of the mean of the differences.

Strictly speaking, the  $r_{12}$  in (25c) should be written as  $r_{M_M}$  so as to indicate that it is a measure of the extent to which successive pairs of means vary together, but it can be shown that the correlation between means is the same as  $r_{12}$ , the correlation between the scores entering into the means.

Since  $M_D = D_M$  and  $\sigma_M = \sigma_{D_M}$  it should be obvious that when testing the null hypothesis we have

$$x/\sigma \text{ (or CR)} = \frac{M_D}{\sigma_{M_D}} = \frac{D_M}{\sigma_{D_M}}$$

That is, the procedure for testing the null hypothesis that  $M_D$  is zero for a population is equivalent to testing the null hypothesis that  $M_{\text{score } 1} = M_{\text{score } 2}$ , where the subscripts 1 and 2 indicate that we are considering 2 populations of scores, 1 for each condition.

Formulas (25a) and (25b) are appropriate in a number of situations in which an  $X_1$  score is somehow paired with an  $X_2$  score. Some of the possibilities are the following:

- $X_1$  as first trial; practice;  $X_2$  as after trial; same person.
- $X_1$  as initial; experience;  $X_2$  as final; same person.
- $X_1$  as pretest; experience;  $X_2$  as posttest; same person.
- $X_1$  under experimental conditions vs.  $X_2$  under normal (or control); same person.
- $X_1$  in one experimental condition vs.  $X_2$  in another; same person.
- $X_1$  as experimental vs.  $X_2$  as control; twin or after pair.
- $X_1$  as experimental vs.  $X_2$  as control; unpaired persons, but matched by pairing on pertinent variables. Data for 2 experimental conditions.

For the last mentioned situation (g), which is commonly employed in experimental work, one can think of having drawn  $N$  individuals at random for one group, then forming the second group by selecting individuals at random, but paired with the members of the first group on the basis of variables which need to be controlled, thus eliminating difference between  $M_1$  and  $M_2$  as<sup>11</sup> not attributable to difference between the 2 groups with respect to the variables used in forming the pairs, since the pairing tends to make the groups equal with respect to the pairing variables. This same pairing procedure (and also twin or after pairs) can be used for

situation  $c$ . Furthermore, as we shall see below, the  $X_1$  and  $X_2$  scores can themselves stand for changes:  $X_1$  the change from pretest to posttest under an experimental condition and  $X_2$  the change under another experimental condition or under control conditions.

The statistical advantages of having scores which are somehow related will be discussed later under the caption "Reduction of sampling errors."

**Difference between independent means.** When we have means for 2 samples which have been drawn independently there will be no way of pairing scores except on a chance basis and chance pairing will tend to produce a zero correlation. In fact, if we took all possible pairs the correlation would be exactly zero. Thus the correlation term in (25b) vanishes, so that the standard error of the difference between means based on independent samples becomes

$$\sigma_{D_M} = \sqrt{\sigma^2_{M_1} + \sigma^2_{M_2}} = \sqrt{\frac{\sigma^2_1}{N_1} + \frac{\sigma^2_2}{N_2}} \quad (26)$$

This formula is not restricted to samples of the same size, i.e.,  $N_1$  need not equal  $N_2$ . The right-hand form of (26) has an obvious computational advantage.

The  $\sigma_{D_M}$  obtainable by formula (26) may be used in exactly the same manner as the standard error of the difference by formulas (25a) and (25b). Again, one sets the null hypothesis that  $M_{\text{true } 1} = M_{\text{true } 2}$  or that the difference between the population means is zero. If it is zero, the sampling distribution of  $D_M$  (resulting from successive replications) will center at zero with  $SD = \sigma_{D_M}$ . If  $D_M/\sigma_{D_M}$  or  $CR$  is sufficiently large one rejects the null hypothesis; if not it is accepted. In other words the general procedure for testing hypotheses about differences is precisely the same for means and other statistical measures as that outlined in the previous chapter. The student would do well to review the discussion dealing with admissible hypotheses, one-tailed vs. two-tailed tests, choice of level of significance, and the 2 types of error one risks in testing hypotheses.

**Differences between other descriptive measures.** The general theory of hypothesis testing is applicable for descriptive measures other than proportions or means. The general pattern for the standard error of the difference between any 2 statistical

measures, say  $S_1$  and  $S_2$ , is

$$\sigma_{D_S} = \sqrt{\sigma^2_{S_1} + \sigma^2_{S_2} - 2r_{S_1S_2}\sigma_{S_1}\sigma_{S_2}}$$

That is, we need to know the standard error for both  $S_1$  and  $S_2$  and a measure of the correlation between  $S_1$  and  $S_2$  in case of non-independence (the  $r$  term drops out for independently drawn samples). The appropriate correlation between means is, as we have indicated, known to be  $r_{12}$ , the correlation between the sets of scores; and the correlation between  $\sigma$ 's is known to be  $r^2_{12}$ ; thus the standard error of the difference between 2  $\sigma$ 's based on the same individuals or on scores related consanguineously or related by pairing on pertinent variables is given by

$$\sigma_{D_\sigma} = \sqrt{\sigma^2_{\sigma_1} + \sigma^2_{\sigma_2} - 2r^2_{12}\sigma_{\sigma_1}\sigma_{\sigma_2}} \quad (27a)$$

and for  $\sigma$ 's based on independent samples

$$\sigma_{D_\sigma} = \sqrt{\sigma^2_{\sigma_1} + \sigma^2_{\sigma_2}} = \sqrt{\frac{\sigma^2_1}{2N_1} + \frac{\sigma^2_2}{2N_2}} = .707\sigma_{D_M} \quad (27b)$$

These formulas are valid for large  $N$ 's (100 or more), and to test the null hypothesis one simply takes  $D_\sigma/\sigma_{D_\sigma}$  as a  $CR$ , with  $\sigma_{D_\sigma}$  being computed by (27a) or by (27b), whichever is appropriate. (For  $N$ 's small, see Chapter 14.)

The difference between medians based on correlated scores cannot be tested because the needed  $r$  is unknown, but for independent samples we have

$$\sigma_{D_{mdn}} = \sqrt{\sigma^2_{mdn_1} + \sigma^2_{mdn_2}}$$

Expressions for  $\sigma_{D_{AD}}$  and for  $\sigma_{D_Q}$  can be similarly written for the case of independent samples.

Any student who is worried because formula (20), on p. 59, for the standard error of the difference between correlated proportions does not include an  $r$  term may rest assured that the correlation has been allowed for even though not visibly so. Formula (20) is analogous to formula (25a), which we have seen is equivalent to the longer formula (25b) in which there is an  $r$ .

### REDUCTION OF SAMPLING ERRORS

One of the aims of scientific method is to attain as great precision in results as is practicable. In statistical work this can be

accomplished by increasing the accuracy or dependability of the scores or individual measurements or responses and by decreasing the chance sampling errors of the various descriptive measures. One way to reduce sampling errors is to employ either the stratified or the area method of sampling, both of which are too complicated for us to discuss here. If the random sampling method is being used in projects which aim to study the difference between groups (or populations), the obvious, and only, way for decreasing the standard error of the difference is to increase  $N$  for either or for both samples. Most field investigations are of this type.

In contrast, the experimentalist can define his population with reference to 2 laboratory or experimental situations, i.e., a population of individuals under situation  $A$  and a population of individuals under situation  $B$ ; his sample individuals for the 2 situations may be the same individuals, first under the  $A$  and then under the  $B$  condition. In general, the use of the same individuals, if feasible in view of possible practice or fatigue effects, will usually involve a fairly high degree of correlation, the net effect of which is to reduce the standard error of the difference considerably; that is, it is sometimes possible to reduce sampling error simply by using the same individuals as the "two" samples. Thus, if we wish to study the effect of 2 different degrees of humidity on mental output or efficiency, it will be a more economical and better controlled experiment if we make observations on the same individuals under the 2 conditions  $A$  and  $B$ , rather than on  $N_1$  individuals under condition  $A$  and  $N_2$  individuals under condition  $B$ .

If it is not feasible to use the same individuals in the 2 experimental situations, we can make up 2 groups by pairing or matching individuals on the basis of 1 or more characteristics. Such a procedure leads to more nearly comparable groups for our experiment than can be obtained by choosing individuals at random and, by using either formula (25a) or (25b) instead of (26), we can make allowance for the fact that the individuals for the 2 samples have not been chosen independently. The use of individuals who have been paired is considered good experimental technique—it cannot be said that a found difference between means for the variable being studied may be due to a lack of comparability of the 2 groups with respect to the matching variables. The use of paired individuals has a statistical as well as experimental advantage in that the sampling error of the difference



between means is thereby reduced without the necessity of increasing the number of cases. If pairing produces an  $r$  of .75, the reduction in  $\sigma_{D_M}$  is equivalent to that achieved by quadrupling the number of cases when the random method of forming groups is employed. After the student has learned about correlation he will better appreciate the fact that the gain in pairing depends upon the extent to which the variables used in pairing are correlated with the variable being studied.

It is thus seen that, for some types of investigations, greater precision can be obtained by judicious planning. If one had unlimited resources, he could always attain any desired degree of precision by simply taking sufficiently large samples.

Frequently the question is raised as to how many cases should be secured for a given study. The answer might be in terms of the number needed to reach a given degree of accuracy, but this in turn would raise the question of what degree of precision is needed, and this in turn depends on how small a difference we wish to detect. When group comparisons are made and when the  $N$ 's are relatively small, the null hypothesis is apt to be accepted too often for the simple reason that a real difference has to be sizable before it is demonstrable by small samples. On the other hand, if a real difference is so small that its statistical demonstration requires thousands of cases, one may question whether it has practical or scientific importance.

### COMPARISON OF CHANGES

Although the comparison of changes involves nothing new in the way of statistical theory, such comparisons are somewhat more complicated than the tests of significance so far discussed. The researcher may be interested in either of 2 questions. First, he may wish to evaluate the effect of only 1 experimental condition or, second, he may wish to contrast the changes produced under 2 (or more) different experimental conditions.

For the first of these, a sample is selected, measurements are made prior to (pretest) and subsequent to (posttest) the provided condition, but, since changes from a first to a second measure might occur because of practice effect or because of some other experience beyond the control of the investigator, it is necessary to set up a control group the members of which are measured and

then remeasured, at chronological times corresponding as closely as possible to those of the pretest and posttest of the experimental group. It is presumed that all uncontrollable effects will be operating similarly on both groups so that any difference in change for the 2 groups will have resulted from whatever was done to the members of the experimental group. The statistical problem is that of evaluating the change shown by the experimental group compared with that shown by the controls.

For the second type of question the investigator starts with 2 experimental groups, one of which is subjected to 1 experimental condition and the other to a second experimental condition, both groups having been measured prior to the experience (pretest), and then again after the experience (posttest). Since the question is concerned with contrasting gains (or losses) associated with the 2 conditions, a control group is not needed. Presumably, uncontrollable factors are alike for the 2 groups. The statistical analysis consists of testing for significance the difference between the changes shown by the 2 groups.

Whether we are dealing with a problem calling for an experimental and a control group or for 2 experimental groups, the 2 groups may be drawn at random or formed on the basis of the pairing of individuals on pertinent variables. If the groups are set up on the basis of pairing we need to allow for that fact when determining the required standard error of the difference between changes.

Parenthetically, it may be said that the setup which involves an experimental and a control group (or 2 experimental groups) for studying shifts has led to a great deal of confusion regarding the proper statistical handling of the data. We have a total of 4 means, for the pretest and the posttest for each of the 2 groups. By using a combination of subscripts, 1 and 2 for the pretest and posttest, and  $E$  and  $C$  to represent the 2 groups, we can specify the means as  $M_{E1}$ ,  $M_{E2}$ ,  $M_{C1}$ , and  $M_{C2}$ . Not all the possible differences between these 4 will have meaning. Those that have meaning may be set forth as:

$D_E = M_{E1} - M_{E2}$ , the change shown by the experimental group.

$D_C = M_{C1} - M_{C2}$ , the change shown by the control group.

$D_1 = M_{E1} - M_{C1}$ , the pretest difference between the groups.

$D_2 = M_{E2} - M_{C2}$ , the posttest difference between the groups.

Which of these 4 meaningful differences should we test for significance? Obviously, it is insufficient to test only  $D_E$  because we can't be sure that the shift shown, even though nonchance, is really due to the interpolated experience. In fact, the reason for having the control group is to enable us to evaluate the shift which takes place as a result of causes other than the experimentally provided experience. Now it might be thought that if  $D_E$  is significant while  $D_C$  is less, or not at all, significant, an effect has been demonstrated. This type of comparison, however, does not provide a check on the net change. Some have argued that if  $D_2$  is significant while  $D_1$  is not, one can safely conclude that the interpolated experience has had an effect. This comparison also fails to test the net change. We should test the significance of the difference between the 2 changes, i.e.,  $D = D_E - D_C$ , in order to gauge properly the net shift. Although, as regards absolute magnitude,  $D_E - D_C$  will always equal  $D_2 - D_1$ , it is easier to evaluate the former difference.

To get the standard error of  $D (= D_E - D_C)$  when the groups have been independently drawn we need the sampling variance of  $D_E$  and  $D_C$  so as to substitute in

$$\sigma_{D_D} = \sqrt{\sigma^2_{D_E} + \sigma^2_{D_C}}$$

Now since  $D_E = M_{E1} - M_{E2}$  is the difference between 2 means based on the same persons, we could get the standard error of  $D_E$  by using formula (25b), but since the difference between correlated means is equal to the mean difference,  $M_{D_E}$ , we can use formula (25a) to get the required  $\sigma^2_{D_E}$ . This same situation holds for the control group, so (25a) would also be used to get  $\sigma^2_{D_C}$ .

If the experimental and control groups have been formed by pairing, our standard error of the difference between changes will require an  $r$  term to enable us to take advantage of the fact that we have a better controlled experiment. The required  $r$  is the correlation between the changes shown by the members of the pairs; to compute it we need to consider the paired changes. We can, however, get the standard error of the difference by way of the algebraic difference between the changes shown by the members of the pairs, without computing an  $r$ .

Let  $X_1$  and  $X_2$  stand for pretest and posttest scores and let the members of the  $J$ th pair be designated as  $J$  and  $J'$ , with  $J$

assigned to the experimental, and  $J'$  to the control, group. Each individual will have a change score which is nothing more than his pretest score minus his posttest score. Thus the change score for the members of the  $J$ 'th pair will be

$$C_j = D_j = X_{1j} - X_{2j} \quad \text{and} \quad C_{j'} = D_{j'} = X_{1j'} - X_{2j'}$$

Hence the difference between the changes (or differences) shown by the members of any pair will be

$$\begin{aligned} D &= (C_j - C_{j'}) = (D_j - D_{j'}) \\ &= (X_{1j} - X_{2j}) - (X_{1j'} - X_{2j'}) \end{aligned}$$

For  $N$  pairs we will have  $N$   $D$ 's. These  $D$ 's are tedious to compute since one must preserve the same direction for each subtraction and keep track of signs. The process can be made somewhat simpler by removing the parentheses, thus

$$D = X_{1j} - X_{2j} - X_{1j'} + X_{2j'}$$

Simply add  $X_{1j}$  and  $X_{2j'}$  and then subtract the sum of  $X_{2j}$  and  $X_{1j'}$  with the sign for  $D$  depending on whether the first or the second of these 2 sums is the larger.

Once the  $N$   $D$ 's have been determined, we can get  $M_D$ ,  $\sigma_D$ , and thence  $\sigma_{M_D}$  by formula (25a). This  $M_D$  will equal  $D_E - D_C$ , or  $(M_{E1} - M_{E2}) - (M_{C1} - M_{C2})$ , and this  $\sigma_{M_D}$  will be exactly the same as

$$\sigma_{DD} = \sqrt{\sigma_{DE}^2 + \sigma_{DC}^2 - 2r_{DEDC}\sigma_{DE}\sigma_{DC}}$$

After the student has learned how to compute  $r$ , he may prefer to use this longer formula for  $\sigma_{DD}$  (equivalent to  $\sigma_{M_D}$ ) rather than go through the tedium of differencing differences. Regardless of how the standard error of the difference is obtained, one tests the null hypothesis by calculating an  $x/\sigma$  (or  $CR$ ), as the net difference between the 2 changes divided by its standard error. The foregoing procedures are also applicable when one is dealing with 2 experimental conditions. One needs only to use appropriate subscripts in place of  $E$  and  $C$ .

The general pattern outlined on pp. 90-92 holds for the comparison of changes for attributes when the groups have been

independently drawn. We merely substitute  $p$ 's (proportions) in place of the  $M$ 's. Thus we would have

$$D_E = p_{E1} - p_{E2} \quad \text{and} \quad D_C = p_{C1} - p_{C2}$$

as changes in proportions for the experimental and control groups. The variance error for  $D_E$  (and also  $D_C$ ) is obtained by formula (20) on the basis of the tabulation scheme set forth on p. 57.

### INFERENCE: ESTIMATION

So far we have discussed statistical inference from the viewpoint of hypothesis testing, but there are occasions when one may wish to use information from a sample as a basis for estimating population values. There are 2 general types of estimation: point and interval. We shall discuss the first briefly in order to introduce some concepts which the student might encounter, and the second because of its practical implications.

**Point estimation.** We may regard a sample statistic as an estimator for the corresponding population value (parameter). How "good" an estimator it is depends on whether or not it is unbiased and consistent, and on its relative efficiency.

An estimator is said to be *unbiased* if the average of a large number of sample estimates tends to equal the parameter being estimated. The mean is unbiased because the mean of sample means will approach nearer and nearer  $M_{pop}$  as we take more and more samples, but  $\sigma^2$  defined as  $\Sigma x^2/N$  is biased in that the mean of sample variances tends to be smaller than the population variance. An unbiased estimate of  $\sigma^2_{pop}$  is given by  $s^2 = \Sigma x^2/(N - 1)$ , but for subtle mathematical reasons  $s$ , or  $\sqrt{\Sigma x^2/(N - 1)}$ , involves a negligible bias as an estimator of the population standard deviation. Note that the bias is small when  $N$  is large.

An estimator is said to be *consistent* if it approaches nearer and nearer the population value as sample size is increased indefinitely. All the measures so far discussed satisfy this criterion.

The *efficiency* of an estimator is a function of its sampling error. Thus, in terms of efficiency the sample mean is far better than the median as an estimator of the central value of a population of normally distributed scores even though both are unbiased and consistent estimators.



**Interval estimation: Confidence interval.** Interval estimation, which takes into account the sampling error of an estimator, provides limits, or an interval, for the population value, and at a prescribed level of confidence. Given a sample mean and its standard error, one could set up a whole series of "trial" hypothesis values for the population mean. All trial hypothesis values well above and below the sample mean could be rejected at a high level (small  $P$ ) of significance, but rejection would become more and more risky as we approached nearer and nearer the sample mean, and for a whole series of values near the sample mean all trial hypotheses would be acceptable. Now this implies that at some point above the sample mean and at some point below the sample mean we change from rejection to acceptance of the trial values. If we have adopted, say, the  $P = .05$  level, the change will obviously be at  $M \pm 1.96\sigma_M$ . In rejecting trial values outside these limits and accepting values within these limits we are in effect inferring that the population value is in an interval defined by these limits.

It would seem that there should be some way of expressing our degree of confidence that the population mean lies between the limits  $M \pm 1.96\sigma_M$ , since, as we have seen, we can be somewhat sure that the sample mean is not a chance deviation from a population mean outside the limits so determined. Note that, given a population mean and sigma, we can legitimately speak of the probability of a sample mean falling in a specified region, but given a sample mean we cannot speak of the probability of the population mean being in a certain region (or interval) for the simple and compelling reason that  $M_{pop}$ , being definitely just 1 value, has no distribution. We can in no way enumerate events so as to conceive of a probability fraction since just 1 event (value) is possible.

In order to arrive at a statement which expresses our degree of confidence, we note that, if we draw a second sample, we would be apt to have a different set of limits for the simple reason that the second sample mean may differ from the first. If we take additional samples of the same size, we would have a distribution of sample means, hence a sort of distribution of sets or pairs of limits, since each sample mean would provide a set. Our discussion can be greatly simplified by taking sets of limits given by  $M \pm 2\sigma_M$  (as approximating the  $M \pm 1.96\sigma_M$  values). For sim-



plicity of exposition, let us assume that we are drawing successive samples from a population having a mean of 10, and that the variability and  $N$  are such that  $\sigma_M$  can be taken as 2. Then  $M \pm 2\sigma_M$  will be  $M \pm 2(2)$ , or  $M \pm 4$ . It will also facilitate our exposition if we think of the random sampling distribution of means in terms of intervals of  $\frac{1}{2}\sigma$  distances on the base line with the approximate percentage area for the several intervals, as shown in the top curve of Fig. 10.

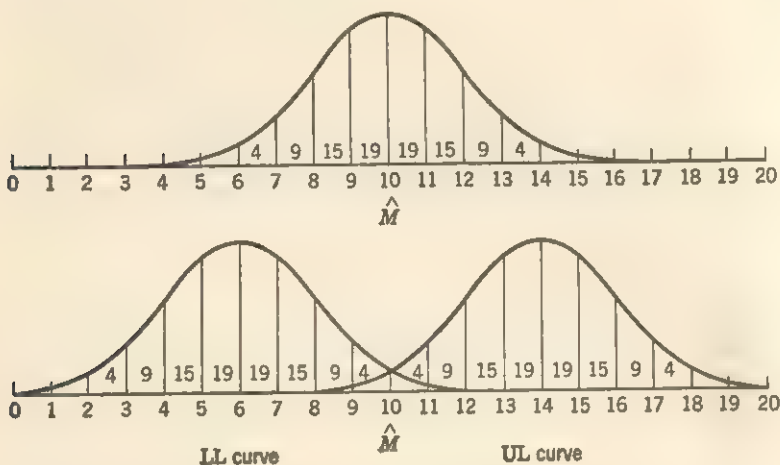


Fig. 10. Generation of confidence limits.

Now each possible sample mean will lead to a lower limit of  $M - 4$  and an upper limit of  $M + 4$ . If we consider the 19 per cent of sample means expected between 9 and 10, we see at once that these 19 will lead to intervals with lower limits between 5 and 6 and upper limits between 13 and 14. That is, the sample means falling between 9 and 10 will generate that part of the lower limit (LL) curve of Fig. 10 between 5 and 6 and that part of the upper limit (UL) curve between 13 and 14. Likewise the 15 per cent of sample means falling between 8 and 9 will lead to the 4 to 5 part of the LL curve and to the 12 to 13 part of the UL curve. Similarly, as can be seen by careful study (a requirement for most students if understanding is to be achieved) of the 3 curves of Fig. 10, every left-hand segment of the top curve generates a left-hand segment for each of the bottom curves. Stated

differently, the left half of the top curve leads to a distribution of intervals with lower limits less than 6 and upper limits of less than 14. In exactly the same fashion it can be seen that the right half of the top curve leads to the right half of the *LL* curve and also the right half of the *UL* curve. Thus we have a sampling distribution of intervals (sets of limits) as found by taking  $M \pm 4$  (or  $M \pm 2\sigma_M$ ). Our next task is to ask how many of these various intervals actually include 10, or the population mean. Reference to Fig. 10 will verify that, out of 100 tries, we would expect to get:

	4	times	an	interval	with	<i>LL</i>	of	2	to	3	and	<i>UL</i>	of	10	to	11
9	"	"	"	"	"	"	"	3	to	4	"	"	"	11	to	12
15	"	"	"	"	"	"	"	4	to	5	"	"	"	12	to	13
19	"	"	"	"	"	"	"	5	to	6	"	"	"	13	to	14
19	"	"	"	"	"	"	"	6	to	7	"	"	"	14	to	15
15	"	"	"	"	"	"	"	7	to	8	"	"	"	15	to	16
9	"	"	"	"	"	"	"	8	to	9	"	"	"	16	to	17
4	"	"	"	"	"	"	"	9	to	10	"	"	"	17	to	18

Notice that for every set of limits in the foregoing groups the population mean *is* in the range or interval defined by the upper and lower limits of the set. When we sum these expected frequencies, we see that 94 per cent of the sets of limits lead to intervals within which the population mean lies. If we had not rounded to the nearest per cent, these would sum to 95.45 per cent. This implies that 4.55 per cent of the time the intervals so defined would not include the population value. This can be verified by noting that sample means of less than 6 (top curve) lead to *upper* limits of *less* than 10, and do so 2.27 per cent of the time, whereas sample means of more than 14 produce *lower* limits of *more* than 10 about 2.27 per cent of the time. These percentages are for the tails of the bottom curves, to the left of the ordinate at 10 for the *UL* curve and to the right of this ordinate for the *LL* curve.

In summary, if one were to make in his lifetime 100 inferences concerning population means on the basis of sample values by each time taking the limits as  $M \pm 2\sigma_M$ , the limits so established would include the population value about 95 per cent of the tries. That is, in the long run he would be correct about 95 per cent of the time in concluding that the population value is within the intervals so determined, and about 5 per cent of the time he would be in error. If he used  $M \pm 1.96\sigma_M$  for setting limits, he would

be correct 95 per cent, and in error 5 per cent, of the time. When we take  $M \pm 1.96\sigma_M$  as a confidence interval, the degree of faith in such limits is represented by a  $P$  of .95; i.e., the *level of confidence* for such an inference is represented by a probability-type figure of .95. If we wish to be surer of our inferences, we might choose the .99 level of confidence, which in practice can be attained by taking  $M \pm 2.58\sigma_M$  as limits.

The limits set by the confidence interval method are so very similar to *fiducial limits*, and the level of confidence, sometimes referred to as the *confidence coefficient*, is so much like *fiducial probability* that the beginning student can well let the mathematical statistician worry about the theoretical difference between what seems to be 2 ways of doing the same thing.

Confidence intervals can be set up for statistical measures other than the mean, but if the random sampling distribution of a given measure is nonnormal the method will not be the simple stunt of taking  $S \pm 1.96\sigma_S$  or  $S \pm 2.58\sigma_S$  where  $S$  stands for any statistical measure. It should be obvious that, since the standard errors for all statistical measures are a function of  $N$ , it is possible by increasing the sample size to narrow the confidence interval without any loss in the degree of confidence with which we accept the limits.

**Confidence interval for a difference.** There are times when it is desirable not only to know whether a difference is significant but also to specify limits for the population difference. Such specification does not presume that a significant difference has been found. Even when a difference fails to reach significance, the specification of confidence limits gives one some idea of the possible difference between population values, and such information may help answer the nonstatistical question of whether the population difference is apt to be large enough to be of practical or scientific importance. This procedure may be helpful in evaluating the consequences of accepting the null hypothesis when the hypothesis is in reality false.

Furthermore, the setting up of a confidence interval may be particularly helpful when we have obtained a difference which is highly significant. Consider the case of a difference of 4.78 inches in mean height between men and their sisters. Because of large  $N$ 's and the presence of brother-sister correlation, the standard error of the difference is very small. Its value is about .07. When

we compute  $D/\sigma_D$  we have a critical ratio of 68. This would, if we could evaluate it, yield a probability, for as large a difference by chance, which would be so microscopically small that we could not comprehend it. However, when we set confidence limits at, say, the .99 level, we have  $4.78 \pm 2.58(.07)$ , or 4.60 and 4.96, as limits for the population difference. This permits a down-to-earth way for evaluating the obtained difference.

**Level of confidence vs. level of significance.** The term "level of confidence" should not, as is frequently the case, be misused in place of "level of significance." The first term pertains to interval estimation, the other to hypothesis testing.

### QUESTION OF ASSUMPTIONS

It may be well to consider briefly the assumptions underlying the procedures so far discussed for making statistical inferences, since assumptions restrict the applicability of a method.

**Independence of sampling units.** It is assumed that the conditions of random sampling hold, but the frequency with which the requirement of independence is violated by researchers suggests that a warning is needed. The violation usually comes about when one makes multiple measurements or observations on each of the individuals in a sample and treats each measurement (or response) as a sample value, thereby inflating  $N$   $n$ -fold times when  $n$  repeated measurements (or responses) are available for each person. The lack of independence comes about in that, for instance, if the sample of individuals happened to include 1 high scoring person there would automatically be  $n$  high scores. The effect of such an inflation of  $N$  is an illegitimate reduction in standard errors.

**Infinite vs. finite universe.** If we are sampling from a finite universe, particularly a universe with a rather small number of cases, it seems reasonable to think that as the sample size becomes large relative to the number of cases in the universe the sample mean, for example, will tend to fluctuate less from the universe mean than is the case when drawing from an infinite population. This suggests that the standard error formulas need to be modified for the finite population situation. The required modifications are available for only a few statistical measures. If we let  $N$  represent the sample size and  $N_{pop}$  the size of the finite universe,

the standard errors for the mean and for a proportion are as follows:

$$\sigma_M = \frac{\sigma}{\sqrt{N}} \sqrt{1 - N/N_{pop}} \quad \text{and} \quad \sigma_p = \sqrt{\frac{pq}{N}} \sqrt{1 - N/N_{pop}}$$

In a given research it is sometimes difficult to decide whether the universe being sampled is finite or infinite in size, and, if finite, it is not always easy to determine the value of  $N_{pop}$ . It might be argued that psychologists never study an infinite universe. It can readily be seen, however, that the corrective factor in the sampling error formulas becomes negligible when  $N_{pop}$  is large. Thus, if  $N_{pop}$  is known to be large relative to  $N$ , it matters little whether the given universe is wrongly conceived as being infinite. For example, when  $N$  is .01 of  $N_{pop}$ , the corrective term leads to a reduction in the sampling error of about .005 of the value obtained by the ordinary formulas.

These formulas for the finite universe situation are frequently useful when we wish to compare a subgroup with a total group which contains the subgroup. Such a comparison is sometimes erroneously made by taking  $\sqrt{\sigma_t^2/N_t + \sigma_s^2/N_s}$  as the standard error of the difference between the subgroup mean,  $M_s$ , and the total mean,  $M_t$ . This makes no allowance for the fact that the 2 means are not based on independent groups. An appropriate procedure is to regard  $M_s$  as based on a sample drawn from a finite universe of  $N_t$  cases with mean and standard deviation of  $M_t$  and  $\sigma_t$ ; then with the standard error of  $M_s$  taken as

$$\sigma_{M_s} = \frac{\sigma_t}{\sqrt{N_s}} \sqrt{1 - N_s/N_t}$$

we can test the significance of the deviation of  $M_s$  from  $M_t$  by using the ratio  $(M_s - M_t)/\sigma_{M_s}$ , which is interpretable as a critical ratio, or  $CR$ . This ratio will give a very close approximation to the  $CR$  which would be obtained if we were to compare the subgroup with the remainder (the total cases less the subgroup) as 2 independent groups, using the usual formula for standard error of the difference. The foregoing scheme would also be applicable in case proportions instead of means were the descriptive measures used as a basis for comparison.

**Skewed distributions.** The standard error formulas given in this chapter assume normal or nearly normal score distributions



for the population being sampled. Skewness is the most frequently encountered evidence for nonnormality, and accordingly it is of interest to consider the effect of skewness on the sampling distribution of the mean, the measure most apt to be involved in testing hypotheses. The relationship between the degree of skewness,  $g_1$ , for a variable and the amount of skewness for the sampling distribution of means is  $g_M = g_1/\sqrt{N}$ . Thus the skewness in the distribution of means rapidly disappears as  $N$  is taken larger and larger. For example, if  $g_1$  is .77 (see Fig. 6, p. 30) and  $N$  is 35, the skewness for the sampling distribution of means will be only .13 (see Fig. 6 again). Accordingly, the procedures in this chapter may be safely used with moderately skewed distributions when  $N$  is large and with markedly skewed distributions when  $N$  is very large. Some methods for handling nonnormal data will be discussed in Chapter 18.

#### A FURTHER WORD ON PROPORTIONS

The student will have noted that the general principles of statistical inference set forth in Chapter 5 have been utilized and extended in the present chapter. There are many points of obvious similarity in the 2 chapters, but there is an additional parallelism which is not obvious. For an attribute involving a dichotomy such as yes-no, like-dislike, pass-fail, etc., we may arbitrarily assign a score of 1 to one category and a score of 0 to the other. That is,  $X = 0$  or 1.

Let  $f_0$  and  $f_1$  stand for the frequency of, say, no and yes responses respectively in a sample of  $N$  cases. Thus we have a miniature frequency distribution, with the 2 categories being analogous to 2 intervals. Let's consider the mean and standard deviation of this miniature frequency distribution, both in terms of gross score formulas. Notice that in Table 9 we have a score

Table 9. SCHEME FOR MEAN AND STANDARD DEVIATION OF A DICHOTOMOUS VARIABLE

Response	$X$	$f$	$fX$	$fX^2$
Yes	1	$f_1$	$f_1(1)$	$f_1(1)^2$
No	0	$f_0$	$f_0(0)$	$f_0(0)^2$
Sums		$N$	$f_1(1)$	$f_1(1)$
			$= \Sigma X$	$= \Sigma X^2$
			$= f_1$	$= f_1$



column,  $X$ , a frequency column,  $f$ , an  $fX$  and an  $fX^2$  column (analogous to  $fd$  and  $fd^2$ , with  $d = X$ ). It will be seen that  $\Sigma X = f_1$ ; hence the mean of the distribution is  $M = \Sigma X/N = f_1/N = p$ , where  $p$  is the proportion of yeses. Hence a proportion may be regarded as a mean.

It will also be seen that  $\Sigma X^2 = f_2$ ; hence when we utilize formula (6b), p. 25, to write the variance of the distribution we have

$$\begin{aligned}\sigma^2 &= \frac{1}{N^2} [N\Sigma X^2 - (\Sigma X)^2] \\ &= \frac{1}{N^2} [Nf_2 - (f_1)^2] \\ &= \left[ \frac{Nf_2}{N^2} - \frac{f_1^2}{N^2} \right] \\ &= (p - p^2) = p(1 - p) = pq\end{aligned}$$

Hence  $\sigma = \sqrt{pq}$  as the standard deviation of the dichotomous distribution. (Any connection with the  $M$  and  $\sigma$  for the binomial?)

In this chapter we have given  $\sigma_M = \sigma/\sqrt{N}$  as the standard error of a mean. If this holds for the dichotomous distribution we would have  $\sigma_M = \sqrt{pq}/\sqrt{N} = \sqrt{pq/N}$ . But this is the same as  $\sigma_p$  given by formula (18b), p. 54, of Chapter 5. This is as it should be since  $p = M$  for the dichotomous distribution.

Furthermore, formula (20) for the standard error of the difference between correlated proportions has its analogue in the development on pp. 83-85 for the difference between correlated means, and formula (21) involves a pattern similar to that of formula (26).

#### NOTE ON THE PROBABLE ERROR

An antiquated procedure is the use of the probable error,  $pe$ , instead of the standard error in connection with sampling. The  $pe$  of the mean is  $.6745\sigma_M$ , and therefore we would expect 50 per cent of successive sample means to fall between  $M_{pop} \pm pe_M$ . Similarly, the  $pe$  for any other statistical measure is .6745 times

its standard error. Since no additional information is yielded by multiplying the standard error by a constant, the continuance of this practice is being discouraged. The student who attempts to survey the research literature on a given topic is apt to encounter *pe*'s and he therefore must know the relationship of the *pe* to the standard error.

## CHAPTER 7

### Small Sample or $t$ Technique

Although the general principles of statistical inference are the same for both large and small samples, the techniques differ. We shall confine our attention in this chapter to the technique for dealing with a single mean and with the difference between 2 means. Chapter 14 will deal with inferences concerning variabilities.

It will be recalled that the sampling distribution of the mean is normal when the trait distribution is normal. This holds regardless of sample size. The sampling distribution of means centers at the population mean with a true standard deviation  $\sigma_M = \sigma_{pop}/\sqrt{N}$ , which sigma we termed the true standard error of the mean. Recall also that the relative deviates,  $(M - M_{pop})/\sigma_M$ , follow the unit normal curve. When successive samples are drawn and a  $\sigma_M$  is computed for each sample by using the sample  $SD$  instead of  $\sigma_{pop}$  (an unknown), the ratios of given  $(M - M_{pop})$ 's to their  $\sigma_M$  values so computed will be distributed normally for very large  $N$ 's and approximately so for  $N$ 's of moderate size, but for  $N$ 's as small as 30 the approximation is none too good. The value 30 is arbitrarily chosen—the approximation to normality becomes progressively worse as we go from large to small  $N$ 's rather than becoming abruptly worse in the vicinity of  $N = 30$ .

We have already mentioned the fact that  $\sigma^2 = \Sigma x^2/N$  suffers from bias, whereas  $s^2 = \Sigma x^2/(N - 1)$  is an unbiased estimator of the population variance. Since the bias in  $\sigma$  increases with a decrease in  $N$ , it is important to use the unbiased estimator when  $N$  is small. We will accordingly use  $s_M = s/\sqrt{N}$ , in place of  $\sigma_M = \sigma/\sqrt{N}$ , as a nonnegligible improvement in the estimate of the standard error of a mean based on a small sample. Even so, the successive sample ratios,  $(M - M_{pop})/s_M$ , with  $s_M$  computed

from each sample, will not follow the unit normal curve because the sampling distribution of  $s$  (also  $\sigma$ ) is skewed for  $N$  small; hence the distribution of successive values of  $s_M$  will be skewed. That is, the successive sample values of  $(M - M_{pop})/s_M$  will involve a variable numerator which is normally distributed and a variable denominator which has a skewed distribution. The distribution of the resulting ratios will be symmetrical about zero but will be less flat-topped than the normal curve, and the smaller the  $N$  the more leptokurtic (less flat-topped) the shape of the curve. Another characteristic of the sampling distribution of  $(M - M_{pop})/s_M$  is that the tails of the curve beyond ratios of about 2 tend to be higher than the tails of the normal curve; that is, there will be relatively more large ratios.

**The  $t$  distribution.** It can be shown that such ratios, involving a normally distributed deviate divided by an unbiased estimate of its sampling error, will follow the so-called  $t$  distribution, defined by

$$y = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right) \sqrt{n\pi}} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}$$

in which  $\Gamma$  indicates the gamma function as defined in texts in advanced calculus. Although this equation will be beyond the mathematical comprehension of most students, it should be noted that  $y$  is the height of a curve, that since  $t$  is squared the distribution is symmetrical, and that the equation contains an  $n$  as yet undefined. This  $n$  has to do with the number of degrees of freedom, a concept which is discussed below. Suffice it to say just now that  $n$  will be a function of sample size (or sizes) and accordingly that there will be not 1 but many distributions of  $t$ , one for each possible value of  $n$ .

Figure 11 shows the curve of  $t$ , when  $n = 7$  and when  $n = 3$ , as compared to the normal curve. For  $n$  larger and larger, the curve of  $t$  approaches that of the normal distribution. Table E of the Appendix gives the values of  $t$ , for  $n$ 's of 1 to 30, which will be exceeded by chance a specified proportion of times. Thus for  $n = 30$  we see from Table E that the  $P = .05$  point is at a  $t$  of 2.04 as compared to a normal deviate of 1.96. For  $n = 10$ ,

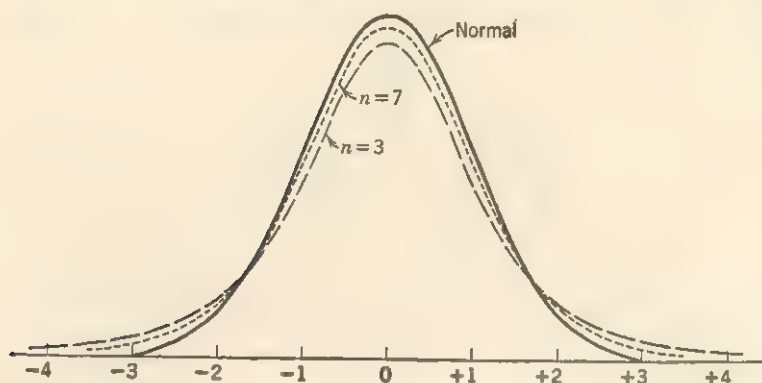


Fig. 11. Normal compared with *t* distribution for  $n = 3$  and  $n = 7$ .

the point corresponding to the .05 level is  $t = 2.23$ . The .01 level is at  $t = 2.75$  for  $n = 30$ , and at 3.17 for  $n = 10$ , as compared with 2.58 for the normal curve.

**Degrees of freedom.** The  $n$  of the equation for  $t$ , and in the  $t$  table, is the number of degrees of freedom ( $df$ ) involved in the estimate of the population variance. The  $df$  depends on how many of the  $x$ 's in  $\Sigma x^2$ , or  $\Sigma(X - M)^2$ , are "free to vary." Suppose two scores, 3 and 5. Their mean is 4, and the sum of squares (of deviations) is  $(3 - 4)^2 + (5 - 4)^2 = 2$ . Now  $\Sigma x = \Sigma(X - M) = \Sigma X - \Sigma M = \Sigma X - NM = \Sigma X - N \frac{\Sigma X}{N} = 0$ , always. There-

fore, as soon as 1 of 2 deviations is known, the other  $x$  is determinable. Thus, if  $x_1$  is  $-1$ , the other deviation,  $x_2$ , must satisfy the equation  $-1 + x_2 = 0$ . One deviation and hence its square can be thought of as dependent upon the other deviation, which has some independence, and therefore 1 degree of freedom. Suppose that we have 3 scores, 3, 4, and  $X$ , which yield a mean of 4. The deviations must satisfy the requisite that they sum to zero; i.e.,  $(3 - 4) + (4 - 4) + (X - 4) = 0$ . Thus 1 of the 3 deviations is fixed by the other 2, i.e., is not independent of their values, because the 3 deviations must sum to zero.

It may be more enlightening to start with symbols for scores. Suppose that  $X_1, X_2, X_3$ , and  $X_4$  represent 4 scores, and it is reported that their mean equals 40. How many of the 4 deviations can we assign at will? Stated in deviation units, we have

$(X_1 - 40) + (X_2 - 40) + (X_3 - 40) + (X_4 - 40)$  as a sum which must equal zero. It is readily apparent that only 3 deviations can "vary freely"—the fourth is fixed by the numerical values of the other 3. Hence  $df = 4 - 1$ ; i.e., 1 degree of freedom in the deviations or their squares is lost because of the 1 restriction imposed. Actually, this restriction comes about because we are taking deviations about 1 constant, the mean, computed from the set of scores at hand. The  $df$  for a sum of squares (of deviations) about a mean is always  $N - 1$  when  $N$  scores are used to compute the mean. In general, the  $df$  for the sum of squares is equal to the number of squares minus the number of restrictions imposed by constants computed from the data.

Note that the unbiased estimate of the population variance,  $s^2 = \Sigma x^2 / (N - 1)$  involves dividing by  $df$ , the number of degrees of freedom. This is a general rule.

**Computation of  $s^2$  or  $s$ .** For  $N$  small the mean and  $s^2$  or  $s$  are readily computed from gross score formulas. Thus  $M = \Sigma X / N$ . To compute  $s^2$  or  $s$  we need  $\Sigma x^2$  in terms of gross scores. This was given earlier (p. 25) as

$$\Sigma x^2 = \frac{1}{N} [N \Sigma X^2 - (\Sigma X)^2] \quad (6a)$$

Dividing this by  $N - 1$  yields  $s^2$ , the square root of which is the required  $s$ . An easily derived relationship between  $s^2$  and  $\sigma^2$  is  $s^2 = \frac{N}{N - 1} \sigma^2$ . Although we do not need a frequency distribution for purpose of computations, a distribution should be made anyway so as to permit at least a rough check on the assumption that the scores have been drawn from a normally distributed population of scores.

***t* for a single mean.** We can test the significance of  $M$  as a deviation from any hypothesized value for the mean,  $M_h$ , by taking  $t = (M - M_h) / s_M$  as an entry in Table E, with  $n = df = N - 1$ , to see whether the obtained  $t$  reaches the  $t$  value required for certain levels of significance. If the  $t$  does not reach the value required for the chosen level of significance, the deviation would be attributed to chance and the hypothesis accepted.

If one wishes to specify the confidence limits for the unknown population mean and to do so with a level of confidence indicated



by  $P = .99$ , he first notes from the table of  $t$  how large  $t$  must be, for the given  $df$ , to correspond to the .01 probability level. Then  $M$  plus and minus the  $t$ , so found, times  $s_M$  will give the desired limits. For example, suppose 9 cases yield a mean of 80 and a sum of squares of 1152. Dividing the sum of squares by  $df$ , or 8, we get  $s^2 = 144$ ,  $s = 12$  as an estimate of  $\sigma_{pop}$ , and  $s_M = 12/\sqrt{9} = 4$ . For 8  $df$  we find from Table E that  $t = 3.355$  for the .01 level. Then  $80 \pm (3.355)(4)$  gives 66.58 and 93.42 as the .99 confidence limits for the population mean. If we used the large sample method of the previous chapter, we would have  $\sigma^2 = 1152/9$ , giving  $\sigma$  as 11.31, from which we would get  $\sigma_M = 11.31/\sqrt{9} = 3.77$ . Since for the normal distribution a relative deviate of 2.575 corresponds to the .01 level, we have  $80 \pm (2.575)(3.77)$  or 70.29 and 89.71 as the .99 confidence limits for the universe mean. These values for the confidence interval differ appreciably from those obtained above when proper allowance was made for the smallness of the sample.

**Difference between correlated means.** It will be recalled that when we have 2 means based on the same individuals or on paired cases, the test of significance of the difference must make allowance for the fact that the 2 sets of scores are not random with respect to each other. In Chapter 6 we saw that this could be done by including the  $r$  term in the standard error of the difference, as in formula (25b), or by working directly with the differences between paired scores. It was shown that  $M_D = D_M$  and that  $\sigma_{M_D} = \sigma_{D_M}$ . When we have small samples, it is easier to work with  $M_D$ , an estimate of the sigma of the distribution of differences between paired scores, and thence  $s_{M_D}$ . To get the best estimate of the sampling error of  $M_D$ , we need the sum of squares of the deviations of the pair differences from the mean difference, i.e.,  $\Sigma(D - M_D)^2$ , which when divided by the proper  $df$ , or  $N - 1$ , where  $N$  is the number of differences or the number of paired scores, gives the best estimate of the variance of the universe distribution of differences. Let  $s_D^2$  stand for this estimate. Then

$$s_{M_D} = \frac{s_D}{\sqrt{N}} \quad (28)$$

The computation is straightforward. Each of the  $D$ 's is the difference between 2 scores, the subtraction being made in the same direction for all, and the sum of squares,  $\Sigma(D - M_D)^2$ , is

obtained by formula (6a) with the  $X$ 's replaced by  $D$ 's; that is  $\Sigma(D - M_D)^2 = \frac{1}{N} [N\Sigma D^2 - (\Sigma D)^2]$ . The  $D$ 's are summed algebraically, and their squares are summed. After  $s_{M_D}$  has been calculated, we get  $t$  as  $M_D/s_{M_D}$ . The hypothesis to be tested is that the universe value of  $M_D$  is zero; the table of  $t$  is entered with the obtained  $t$  and with  $df = N - 1$  in order to see whether it reaches a prescribed level of significance. Note that the  $df$  is 1 less than the number of  $D$ 's, not 1 less than the total number of scores (see "Further note" on  $df$ 's, p. 111).

The assumption of normality pertains to the  $D$ 's; hence, again, even though a frequency distribution is not needed for computational purposes, it should be made so as to provide a rough check on the assumption. A confidence interval for  $M_D$  (and consequently  $D_M$ ) can be set up in precisely the same manner as indicated above for a single mean.

**Difference between independent means.** Given 2 groups of  $N_1$  and  $N_2$  cases, and that we wish to test the significance of the difference,  $D_M = M_1 - M_2$ . By the procedure of the previous chapter for large  $N$ 's, we would make the necessary calculations for determining  $D_M/\sigma_{D_M}$  or  $CR$ . As an aid to transition in thought from  $CR$  to  $t$ , let us first write the expression for  $CR$ , thus

$$CR = \frac{D_M}{\sigma_{D_M}} = \frac{M_1 - M_2}{\sqrt{\sigma^2_{M_1} + \sigma^2_{M_2}}} = \frac{M_1 - M_2}{\sqrt{\frac{\sigma^2_1}{N_1} + \frac{\sigma^2_2}{N_2}}}$$

which involves the 2 sample variances. Now, for the small sample situation, we need  $t = D_M/s_{D_M}$  where  $s_{D_M}$  is to be the best possible estimate of the standard error of the difference. To get this we apparently need the best possible estimates of the 2 variances of the 2 populations from which the samples have been drawn. But here we encounter an assumption underlying  $t$  for this situation: the 2 populations must have the same variance. Hence, we need just 1 estimate, an estimate of the variance common to the 2 populations. Calling this estimate  $s^2$ , by analogy with the  $CR$  technique, we need

$$t = \frac{D_M}{s_{D_M}} = \frac{M_1 - M_2}{\sqrt{\frac{s^2}{N_1} + \frac{s^2}{N_2}}}$$

The best estimate,  $s^2$ , of the common population variance is obtained by computing the sum of squares separately for the 2 samples, then combining these sums, and dividing by the proper *df*, or

$$s^2 = \frac{\Sigma(X - M_1)^2 + \Sigma(X - M_2)^2}{N_1 + N_2 - 2}$$

The 2 separate sums are computed by formula (6a). Note that 2 degrees of freedom are lost because the sum of squares is about 2 means, which leads to 2 restrictions. Substituting the obtained  $s^2$  in the above expression leads to a *t*, which is looked up in Table E with *df*, or *n*, equal to  $N_1 + N_2 - 2$  in order to see whether it reaches a chosen level of significance.

There is one point in the method of determining the  $s^2$ , needed for testing the significance of the difference between means, which may have puzzled the student. The setting of the null hypothesis, in combination with the assumption of equal population variances, implies that the 2 samples have been drawn from a single universe or from 2 universes which have the same mean and equal variances, for the given and measured trait. It might accordingly be assumed that the best estimate of the population variance would be obtained by taking the sum of squares about the combined mean rather than about the separate means. The former would give the better estimate of the variance if it were actually known that the 2 universe means were the same (or that only 1 universe was involved), but there is always the possibility that the 2 universe means really differ; if this were true, the taking of the sum of squares about the combined mean would, in general, yield too large an  $s^2$  for the simple reason that the real difference between groups would be contributing to the variability of the 2 groups combined. (The student who has difficulty seeing this point should imagine what would happen to the variance of scores when 2 groups markedly different in means were combined.) It follows, therefore, that in the long run the best value for  $s^2$  will be provided by summing the sums of squares about the 2 means.

The procedure for setting a confidence interval when we have independent means is no different from that for correlated means. Simply take  $D_M \pm t_\alpha s_{D_M}$  where  $t_\alpha$  is the *t*, for the given *df*, required for significance at the  $P = \alpha$  level. This will give limits for the  $P = 1 - \alpha$  level of confidence. Suppose we wish the .99

confidence interval; this requires an  $\alpha$  of .01, or as sometimes written,  $t_\alpha = t_{01}$  where  $t_{01}$  is found under the  $P = .01$  column, opposite the  $df$ .

**Further note on degrees of freedom.** Suppose 2 independent groups with  $N_1 = N_2 = N$ , and also 2 groups of scores based on  $N$  cases (or  $N$  paired persons). For the former the  $df$  is  $N_1 + N_2 - 2 = 2N - 2$ , whereas for the latter the  $df$  is  $N - 1$  even though in the paired situation the total number of persons is  $2N$ . This may be (and has been) confusing to some; it seems as though the obviously better plan (matching) leads to a loss in  $df$  compared to the setup involving independent groups. It is sometimes argued that the  $df$  would perhaps be larger if we worked not with the difference scores but with the 2 sets of scores in terms of the sums of squares of deviations for each set and the sum of cross products since, as can be seen from p. 84,

$$\Sigma(D - M_D)^2 = \Sigma x_1^2 + \Sigma x_2^2 - 2\Sigma x_1 x_2$$

The  $df$  for the left-hand sum of squares is obviously  $N - 1$ , and since the right-hand side of the equation is merely an algebraic variant of the left-hand side, it does not seem reasonable to believe that the  $df$ 's will differ for the 2 sides. Note that if we consider  $\Sigma x_1^2$  as having  $N - 1$  degrees of freedom, we cannot have any more degrees of freedom for the other sums on the right side because the  $x_2$  values are not independent of the  $x_1$  values; they (the  $x_2$  scores) are not "free to vary."

**Comparison of changes.** In the last chapter (p. 90) we discussed the procedures for testing the differences between changes shown by 2 groups. For the situation involving paired persons, a  $D$  for the difference between changes for the members of a pair was defined (p. 93), and the test of significance involved computing, for  $D$ 's so defined, an  $M_D$ ,  $\sigma_D$ , and thence  $\sigma_{M_D}$ . For the small sample, or  $t$ , technique we need  $s_D$  and  $s_{M_D}$ , just as given above for correlated means. The  $df$  is 1 less than the number of pairs. For the setup involving the changes for independent groups, we would need an  $s_{D_D}$  instead of the  $\sigma_{D_D}$  of p. 92. The required  $s_{D_D}$  is given by

$$s_{D_D} = \sqrt{\frac{s_D^2}{N_E} + \frac{s_D^2}{N_C}}$$

in which

$$s^2_D = \frac{\Sigma(D - M_{DE})^2 + \Sigma(D - M_{DC})^2}{N_E + N_C - 2}$$

with the subscripts  $E$  and  $C$  referring to experimental and control groups. Thus, the procedure for testing hypotheses involving changes for 2 groups is precisely the same as that for testing the difference between 2 independent means, discussed above.  $X$  is replaced by  $D$ , a difference score.

**One-tailed versus two-tailed test.** Our discussion of the  $t$  technique so far has been in terms of the  $t$  value needed for a two-tailed test at a given level of significance. If the hypothesis to be tested or the decision to be made logically warrants a one-tailed test the  $t$  required for significance at the .01 level would be found under the .02 column of Table E, and for the .05 level the .10 column would be used. Those who do not wish to be restricted to the  $P$  levels given in Table E will find for  $df$ 's up to 20 the  $P$  associated with any  $t$  in Table XLV of Peters and Van Voorhis' *Statistical procedures and their mathematical bases*. This table gives one-tailed values, which need, of course, to be doubled for two-tailed tests.

**Some comments and cautions.** It might be thought that the assumption of normality underlying the use of  $t$  could be tested on the basis of the sample (or samples) at hand either by testing the departure of  $g_1$  (skewness) and  $g_2$  (kurtosis) from zero (or by a chi square technique to be discussed in Chapter 13), but these methods of testing for normality are not sensitive enough to lead one to reject, on the basis of a small sample, the hypothesis of normality unless the departure therefrom is very marked. Likewise, the as yet undiscussed test (see Chapter 14) for a possible difference between variances is too insensitive when used with small samples to lead to rejection of the hypothesis of equal variances unless the difference between the 2 universe variances is sizable; hence it is difficult to be sure that the assumption of equality of variances is tenable when 2 groups are being compared by the  $t$  technique. The foregoing statements are, of course, based on the proposition that by statistical methods one can prove, at a desired level of significance, that a sample distribution did *not* arise from a normally distributed universe or that 2 universe values are different, but such methods will not prove normality nor prove that 2 universe values are identical.



A method for testing the significance of the difference between 2 means when the assumption of equality of variances is not tenable may be found in section 4.14 of *Experimental designs* by Cochran and Cox. Some methods for handling nonnormal data are given later (Chapter 18).

Suppose that in 1 study the difference between 2 means for 2 small samples leads to a  $t$  which falls at the .01 level and that in another study 2 large samples yield means, for another trait, which are also significantly different at the .01 level. Can we place as much reliance on the first difference as on the second? The answer is yes, provided the 2 studies have been carried out with the same degree of care as regards controls and adequate sampling techniques, and provided it is safe to presume that the fundamental assumptions underlying  $t$  are tenable. Thus our confidence in a result based on small samples is a function not only of the probability level of significance attained but also of our faith that assumptions have been met. Since, as we have suggested, the conditions of trait normality and equality of variances are exceedingly difficult to demonstrate when the only information available is based on the small samples at hand, we are forced to conclude that, in general, we cannot place as much reliance on the results from small samples as on those from large samples.

This raises the question of the place of small samples in psychological research, and about this there will be a diversity of opinion. We do not propose to settle the issue or even debate it; instead, we shall mention a few points which we feel are pertinent. There are, of course, types of research for which it is impossible or practically impossible to secure more than a few cases either because of their scarcity or because of prohibitive costs. For such situations it is fortunate that the small sample or  $t$  technique, which permits some allowance for the smallness of the sample or samples, is available. Quite frequently small samples may be useful in a preliminary study which is carried out solely for the purpose of guiding the experimenter. If given hypotheses seem to be verified, then the next step should be to secure more cases for further verification rather than to rush into print with positive conclusions.

It seems to the writer that those who publish statistical results based on a small number of cases should, unless they are positively sure that the basic assumptions underlying  $t$  have been met



1. The first part of the report is a general statement of the purpose of the study. This is followed by a brief review of the literature on the subject. The next section is a description of the methods used in the study. This is followed by a presentation of the results of the study. The final section is a discussion of the results and their implications.

# THE HISTORY OF THE UNITED STATES OF AMERICA

The history of the United States of America is a story of a young nation that grew from a small colony of settlers to a great power. The story begins with the first settlers who came to the New World in search of a better life. They found a land of opportunity, but also a land of conflict. The settlers fought with the Native Americans for land and resources. They also fought among themselves for power and influence. The story continues with the American Revolution, which was a struggle for independence from British rule. The revolution was a success, and the United States became a new nation. The story then moves on to the period of westward expansion, which was a time of great growth and discovery. The settlers moved westward, seeking new lands and new opportunities. They found a land of great beauty and great resources, but also a land of great conflict. The settlers fought with the Native Americans for land and resources. They also fought among themselves for power and influence. The story ends with the present day, where the United States is a great power and a land of great opportunity.

The history of the United States of America is a story of a young nation that grew from a small colony of settlers to a great power. The story begins with the first settlers who came to the New World in search of a better life. They found a land of opportunity, but also a land of conflict. The settlers fought with the Native Americans for land and resources. They also fought among themselves for power and influence. The story continues with the American Revolution, which was a struggle for independence from British rule. The revolution was a success, and the United States became a new nation. The story then moves on to the period of westward expansion, which was a time of great growth and discovery. The settlers moved westward, seeking new lands and new opportunities. They found a land of great beauty and great resources, but also a land of great conflict. The settlers fought with the Native Americans for land and resources. They also fought among themselves for power and influence. The story ends with the present day, where the United States is a great power and a land of great opportunity.

standard deviations, permit us to write the linear equation for predicting  $Y$  from  $X$  or  $X$  from  $Y$ .

Our present discussion will be concerned with the determination of relationship between such typical variables as height, weight, strength, age, intelligence, social status, attitudes—i.e., with those variables which show variation from individual to individual. The question of the relationship between variables of this type can be stated quite simply: Is there a tendency for the individual who ranks high (or low) on one characteristic to be high (or low) on another also? It should be noted that at times a relationship may involve just 1 variable: Are heights of sons related to the heights of their fathers? Are the IQ's of adults related to their childhood IQ's?

### THE SCATTER DIAGRAM

The first task is that of tabulation. If we have observations on the height and weight of a large number of individuals, using cross-sectional or coordinate paper, we can lay off on the  $y$  axis convenient tabulating intervals for, say, height and on the  $x$  axis intervals for weight. The rules for choosing intervals stated on p. 6 should be followed here. Tabulation then consists first of finding on the  $y$  axis the interval in which an individual's height falls and locating the interval on the  $x$  axis for his weight. A tally or dot is then placed in the *cell* formed by the intersection of these 2 intervals. The result of such a two-way or cross tabulation is referred to as a *scatter diagram* or correlation table. It will contain as many tallies as there are pairs of observations. The tallies in each row, or horizontal array, can be counted and recorded, separately by rows, to the right of the diagram. This procedure will, of course, yield the frequency distribution for all individuals with respect to the variable on the  $y$  axis. A similar count, and recording at the top, of tallies for each column, or vertical array, will yield the distribution for the other variable. The sum of the frequencies for either of these marginal distributions should equal  $N$ , or the number of pairs of observations.

Figures 12a and 12b are illustrative scatter diagrams, but not models so far as number of grouping intervals is concerned. In practice, from 12 to 20 intervals should be used in order to reduce the grouping error to a negligible amount. It is to be understood

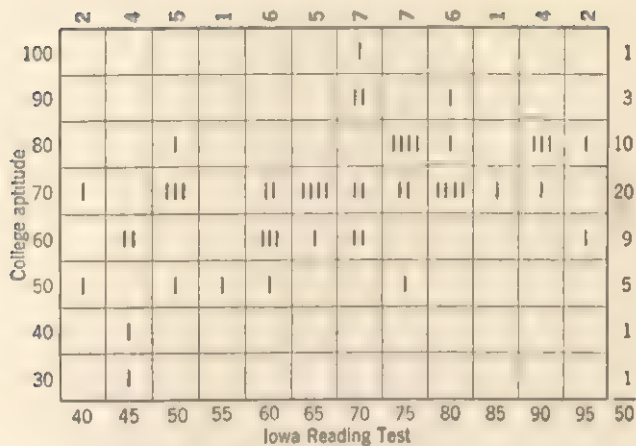


Fig. 12a. Correlation scatter diagram for 2 tests

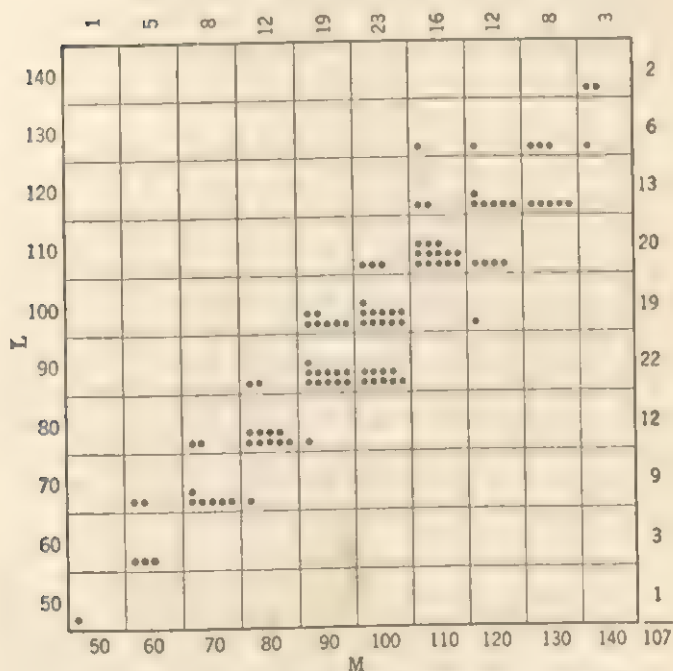


Fig. 12b. Correlation scatter for 2 forms of Stanford-Binet.

that the intervals in these charts are 40-44, 30-39, 50-59, etc. The student should study these diagrams so as to grasp some of the mechanical details involved in their construction. It should be noted that the number and size of the intervals for the 2 variables need not be the same, and that the zero points on the scales of measurement need not appear or even be indicated on the axes.

It can readily be seen that these 2 diagrams represent different degrees of relationship. A precise method for measuring or describing degree of relationship or association or correlation will be discussed in detail in the pages to follow. We shall begin with a symbolic definition of a basic correlation coefficient, indicate its computation, and then discuss its meaning, interpretation, assumptions, and finally its limitations. Certain elementary mathematical derivations will be either indicated or given whenever it is thought that their inclusion will be useful in clarifying a point or clinching an assumption.

The Pearson *product moment correlation coefficient* is defined by

$$r = \frac{\sum xy}{N\sigma_x\sigma_y} \quad (29)$$

in which  $x$  and  $y$  represent deviation measures from the respective means of the 2 variables, i.e.,  $x = X - M_x$  and  $y = Y - M_y$ , the sigmas in the denominator are the standard deviations of the 2 distributions, and  $N$  is the number of individuals measured. With reference to a scatter diagram,  $M_x$  and  $\sigma_x$  hold for the marginal distribution at the top, whereas  $M_y$  and  $\sigma_y$  hold for the distribution to the right. The numerator term,  $\sum xy$ , implies that the product of each individual's  $x$  and  $y$  is determined, and that all such products are summed algebraically. There will, of course, be  $N$  products in this sum, some of which will be positive, some negative, and perhaps some zero.

Definition formula (29) is seldom used for computation. For  $N$  small a usable computational equivalent is

$$r = \frac{N\sum XY - \sum X\sum Y}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}} \quad (30)$$

which involves 4 familiar sums, and the sum of the products of the paired raw scores. This formula is unwieldy for large  $N$  and or

scores which are numerically large. For reasons which will become apparent later, the careful researcher will always make a scatter diagram, and once this has been done it is economical to compute  $r$  in terms of step-interval deviations from arbitrary origins. An appropriate formula is

$$r = \frac{N\sum d_x d_y - \sum d_x \sum d_y}{\sqrt{N\sum d_x^2 - (\sum d_x)^2} \sqrt{N\sum d_y^2 - (\sum d_y)^2}} \quad (31)$$

in which  $d_x$  is defined as an individual's score deviation, in step intervals, from an arbitrary origin on the  $X$  scale, and  $d_y$  is defined similarly for the  $Y$  scale. The student will note the similarity of the radical terms to formula (5) for computing  $\sigma$ . Formula (31) calls for 2 sums, 2 sums of squares, and a sum of cross products, all in terms of step or interval deviations from arbitrary origins. The arbitrary origins may be taken at the center or at the bottom of each distribution. The former will involve handling smaller figures but will have the disadvantage of introducing negative numbers. The latter scheme is better if a calculating machine is available.

#### CALCULATION OF $r$

The computation of  $r$  will be illustrated for both hand and machine calculating methods. The hand calculation scheme here used may not be quite as economical as other available schemes, but the particular setup has the advantage that it forms an economical basis for machine computation, and the author presumes that practically all those who are apt to compute more than a few  $r$ 's will have access to a calculating machine of the Monroe or Marchant or Friden type. Once the steps involved in the hand calculation form are grasped, it becomes easy to transfer them to machine work. The writer has never found the commercial correlation charts helpful. All one needs is a sheet of cross-section paper ruled 4 lines to the inch, on which one can readily lay out the axes, in intervals, for tabulating or tallying. When the scatter diagram has been made and the tally (or dot) marks have been summed across and up to get the marginal frequencies (as shown in Figs. 12a and 12b), the  $d$  values, taken from an arbitrary origin at the bottom-most interval for each variable, can be written, preferably with colored lead, alongside the marginal frequencies



(see Table 10). The columns of  $fd$  and  $fd^2$  values along each margin can be obtained by multiplying in exactly the same manner as was previously done for calculating the standard deviation. The sums of these columns provide 4 of the 5 sums needed for  $r$ .

Table 10. \* COMPUTATION OF  $r$ 

$f$	$d_x$	$fd_x$	$fd_x^2$	$d_y$	$fd_y$	$fd_y^2$	$d_x d_y$
110	1	1	1	3	3	9	3
100	2	2	4	4	8	16	8
90	3	3	9	5	15	25	15
80	4	4	16	6	24	36	24
70	5	5	25	7	35	49	35
60	6	6	36	8	48	64	48
50	7	7	49	9	63	81	63
40	8	8	64	10	80	100	80
30	9	9	81	11	99	121	99
20	10	10	100	12	120	144	120
10	11	11	121	13	143	169	143
Σ	61	224	253	61	224	253	224
$\Sigma fd_x$		224		$\Sigma fd_y$		224	
$\Sigma fd_x^2$			253	$\Sigma fd_y^2$		253	
$\Sigma d_x d_y$							224

$$(61)(1007) - (224)(253)$$

$$r = \frac{(61)(1007) - (224)(253)}{\sqrt{61(1012) - (224)^2} \sqrt{61(1207) - (253)^2}} = .776$$

\* Space limitations prevented the use of two few intervals in this table. A complete table of  $d_x$  and  $d_y$  values would be 25-29 on the  $x$  axis and 60-64 on the  $y$  axis.

In order to obtain  $\Sigma fd_y$ , each individual's  $d_y$  must be multiplied by his  $d_x$  and all such products then summed. In the 110 interval on the  $y$  axis we find 1 individual whose score on the  $X$  variable falls in the 50 interval on the  $x$  axis. In terms of step deviations his  $d_x$  value is 8 and his  $d_y$  value is 3 and therefore 3 times 8, or 24, represents his  $d_x d_y$  product. Another individual with the same  $d_y$  value has a  $d_x$  value of 6, whence 6 times 8 is his contri-

bution to  $\Sigma d_i d_j$ . The third individual in the 140 interval has a  $d_i$  value of 7, whence 7 times 8 is his product. These 3 individuals contribute  $5 \times 8 + 6 \times 8 + 7 \times 8$ , or 144, to the sum of products. The  $d_j$  value of 8 is a common factor to these 3 products, whence  $8 \times (5 + 6 + 7)$  or  $8 \times 18$  yields 144. This suggests a scheme for computing the  $d_i d_j$  sum, which involves first summing the  $d_i$  values for a particular  $Y$  interval or array and then multiplying this sum by the  $d_j$  value. Thus the  $d_i$  values of the 3 individuals in the 130 interval sum to 34, and in the 120 interval to 34, and so on down to the 60 interval, which yields 2 as the sum of the  $d_i$  values. The determination of these  $d_i$  sums is greatly facilitated by the use of a runner on which the  $d_i$  values 0, 1, 2, 3, . . . , have been labeled to correspond exactly with the deviations in step intervals alongside the marginal distribution at the top of the diagram. Since each of these  $d_i$  sums is to be multiplied by a  $d_j$  value and then all the products summed, it is convenient first to record the  $d_i$  sums to the right as a separate column and then to multiply each  $d_i$  sum by the corresponding  $d_j$  value, thus leading to the last column of figures. Before these five multiplications are made the column of  $d_i$  sums should be added to see whether it agrees with the  $\Sigma d_i$  already computed from the marginal distribution of  $X$  scores. Thus an internal check is provided for the column of  $d_i$  sums; all other computations should be done twice in order to insure accuracy.

When a calculator is available, the work sheet need not include the  $df$  and  $df'$  columns, since the sums of these 2 columns can readily be obtained by the method discussed on pp. 23-24. This means that the column of  $d_i$  sums can be placed alongside the  $d_j$  values; then each  $d_i$  sum can be multiplied by the juxtaposed  $d_j$  value, with the products allowed to accumulate in the dial as the needed  $\Sigma d_i d_j$ . Thus the right-hand column figures need not appear on the work sheet.

The substitution of the 5 sums into formula (31) is straightforward. The denominator factors are evaluated as explained on p. 24, and the numerator is obtained by punching  $\Sigma d_i d_j$  into the keyboard and multiplying by  $N$ , then with the product left in the lower dial,  $\Sigma d_i$  is subtracted  $\Sigma d_j$  times. If needed, the 2 means can be obtained by substituting  $\Sigma d_i$  and  $\Sigma d_j$  into (3), and the 2 standard deviations by multiplying the proper radical by the interval size and dividing by  $N$  (equivalent to substituting the sum and sum of squares into (5)).

## CHAPTER 9

### Correlation: Interpretations and Assumptions

Intelligent use of the correlation coefficient and critical understanding of its use by others are impossible without knowledge of its properties. It is not sufficient that we be able merely to recognize  $r$  as a measure of relationship. It is a peculiar kind of measure which permits certain interpretations provided certain assumptions are tenable and provided one considers possible disturbing factors. Since the interpretations of  $r$  are so closely related to assumptions, no attempt will be made to present a separate discussion of these 2 aspects. The factors which affect  $r$ , and which are therefore limitations additional to assumptions, will be discussed in Chapter 10.

#### STUDY OF SCATTERGRAM

We shall begin by making a somewhat detailed study of certain properties of a typical scatter diagram. The columns and rows of the diagram have already been referred to as vertical and horizontal *arrays*, the intersection of 2 arrays has been called a *cell*, and the meaning of the marginal distributions has been given. If the scatter diagram depicted in Table 11 is examined, it will be noted that each vertical (and also each horizontal) array contains a frequency distribution, and that the marginal totals really represent the number of cases in these array distributions. These array distributions are very much like any other typical distribution: bell-shaped with a clustering or scattering about a central value. The mean and standard deviation again become useful descriptive terms. Thus, in Table 11, the mean height of sons whose fathers were 64 inches tall is found to be 66.8 inches. This is simply the mean of the 12 cases which fall in this particular array. Similarly for all the vertical arrays we have the means as

recorded along the bottom of Table 11. The means of the horizontal array distributions have been recorded to the right of the scatter diagram. For example, the mean height of the 10 fathers whose sons were 72 inches tall is 70.0 inches.

Table 11. CORRELATION TABLE FOR HEIGHT OF FATHERS (X) AND HEIGHT OF SONS (Y)

	2	6	12	19	27	26	20	26	20	15	8	5	$M_x$
75										1			1 71.0
74										1		1	2 72.0
73							1	1		1	1	1	5 70.6
72						1	1	2	2	2	1	1	10 70.0
71				1	2	2	2	3	4	2	2	1	19 69.1
70			1	1	4	2	4	4	4	3	1	1	25 68.5
69		1	1	3	4	3	5	6	4	2	1		30 67.8
68		1	2	2	5	6	5	6	3	2	2		31 67.7
67	1	1	3	4	5	5	4	2	3	1			20 66.7
66		1	2	2	2	4	3	1					15 66.3
65		1	2	3	2	2	1	1					12 65.8
64	1	1	1	2	2	1							8 64.7
63				1	1								2 65.5
	62	63	64	65	66	67	68	69	70	71	72	73	
$M_y$	65.5	66.5	66.8	66.8	67.6	67.8	68.6	69.1	69.5	70.6	70.3	72.0	
$r = .56$													$N = 192$
$M_x = 67.00$													$\sigma_x = 2.19$
$M_y = 68.44$													$\sigma_y = 2.33$

If the means of the vertical arrays are plotted (see crosses in Fig. 13) two things will be noticed: the means are progressively greater as we pass from short to tall fathers, and they fall approxi-

mately on a straight line. It will be noted (see dots in Fig. 13) that the means for the horizontal arrays also approximate a line and show progression. Now, with reference to the means of the vertical arrays, each represents the mean height of sons of fathers of a particular height and therefore may be used as a basis for predicting the height, if unknown, of a man if we have been told

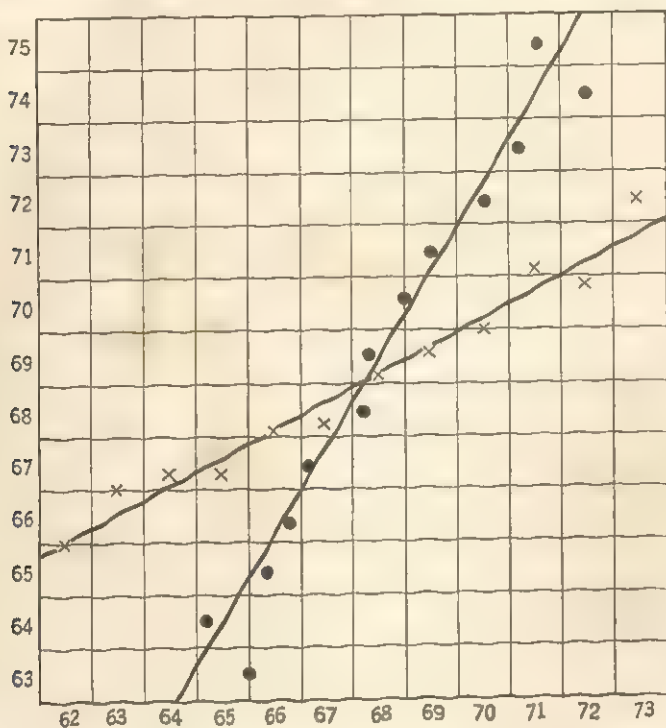


Fig. 13. Plot of array means for data of Table 11.

the height of his father. Thus, if the father is 66 inches tall, the best estimate of his son's height is 67.6, the observed mean height of men whose fathers are 66 inches in height.

Obviously such an estimation would be subject to considerable error, since we have also the observable fact that the heights of sons of fathers 66 inches tall show a large amount of variation about the array average. This variation tells us something about the possible magnitude of the error involved in using 67.6, the array mean, as our estimated value. The unknown height, of

which we take an array mean as an estimate, may actually fall anywhere within rather wide limits on either side of the array mean. These limits can be described in terms of the standard deviation of the array distribution; i.e., the error of estimate can be stated in terms of a  $\sigma$ . The standard deviation for the distribution of heights of sons whose fathers were 66 inches in height is about 2.1. Now, if we take 67.6 as the best estimate, we can say that, if we were to predict the height of 100 sons (fathers 66 inches), about 68 per cent of the time the error would be within the limits  $67.6 \pm 2.1$ , 95 per cent within  $67.6 \pm 4.2$ , and nearly always within the limits  $67.6 \pm 6.3$ . Likewise, when the sigmas for the several arrays have been computed, a statement of the limits of the error in predicting any son's height from his father's height can be made. Such a procedure will yield as many measures of error as there are vertical arrays. We shall soon see that a convenient assumption can be made which will usually allow us to use a single indication of the error of estimate.

Let us return again to the line of the means. Two such lines have been drawn in Fig. 13; one line "fits" the means of the vertical, the other the means of the horizontal, arrays. Let us for the present confine our attention to the means of the vertical arrays. They do not lie exactly on the drawn line; some are above, some below. If they fell exactly on the line, a prediction based on an array mean would be precisely the same as a prediction obtained by noting the  $Y$  value of the line where it cuts the middle of the array. Furthermore, if the means were exactly on a straight line, we might write the equation for this line in the form  $Y = BX + A$ , where  $A$  equals the  $y$  intercept (value of  $Y$  where line crosses the  $y$  axis) and  $B$  equals the slope of the line (the inclination of the line to the  $x$  axis). With  $A$  and  $B$  known, the value of  $Y$  for a particular  $X$  can be readily estimated.

But, since the means do not lie exactly on a straight line, the above reasoning would not seem offhand to yield us anything of practical value. From many viewpoints, however, it is desirable that we determine the equation of the straight line which best "fits" the means, i.e., the equation of a line which passes near all the means. Then we can use this equation instead of the array means in making predictions. The justification for this procedure depends upon the validity or tenability of an assumption: we assume that the failure of the means to fall exactly on a straight



line is due to chance fluctuations in the means. Each array mean is based on a sample and consequently deviates more or less from the true or population value of the mean for the array. This is equivalent to saying that, if all the array means were based on a much larger number of cases, we could assume that they would approximate more exactly a straight line. This is an assumption which can always be made provided the array means for a particular scatter do not show marked deviations from linear form. (Adequate checks in terms of probability, to be described later, can be utilized to ascertain whether the fluctuations from linearity are larger than is reasonable on the basis of chance.)

### THE BEST-FIT LINE

We can now consider one of the advantages of using a line instead of the several array means as a basis for prediction. The location of the line is dependent upon all the means, or rather upon all the cases. It therefore seems reasonable to believe that the line would be more stable from the sampling viewpoint than would the array means, each of which is based on a rather small number of cases.

If we accept the assumption of *linearity* of array means, our problem is that of determining  $A$  and  $B$  so that we can write the equation of the line of means. We need the equations of two lines:  $Y = BX + A$  for the means of the vertical arrays and  $X = B'Y + A'$  for the horizontal array means. We shall consider the determination of the constants  $A$  and  $B$  for the first equation, but before doing so something must be said concerning what is meant by a "best-fit" line. The constant  $A$  gives the  $y$  intercept, i.e., tells us where the line cuts the  $y$  axis. Suppose we think of several possible lines having the same slope (the same  $B$ ) as the line in Fig. 13 which passes near the crosses. Obviously, if we considered a line passing near the top or bottom of the scatter diagram, it would be a "worse fit" than that drawn in Fig. 13. Likewise, if we think of pivoting the line about some point, thereby altering its slope, it can be readily seen that rotating it to a vertical or horizontal position would give a worse fit. It should now be clear that the assigning of some values to  $A$  and  $B$  will lead to a worse fit than that obtained by certain other values, or conversely that some values will yield a much better fit than others.

One criterion which is accepted as a basis for a best-fit line is that the sum of the squares of the deviations from the line shall be as small as possible. With respect to determining the best-fit line to the means of the vertical arrays, this criterion or definition of fit implies that the values of  $A$  and  $B$  are to be such that the sum of the squared deviations of the observed heights of sons—deviations in an up and down or vertical direction—about the line will be a minimum. Stated in symbols, let  $Y' = BX + A$ , where  $Y'$  (read  $Y$  prime) is the value estimated from a given  $X$ , and let  $Y$  be the observed value. Then  $(Y - Y')^2$  represents the squared deviation of any  $Y$  from the line or estimated value. The problem is so to choose  $A$  and  $B$  as to make  $\Sigma(Y - Y')^2$  as small as possible. It is more convenient to deal with both the equation,  $y' = bx + a$ , and the sum,  $\Sigma(y - y')^2$ , in deviation units, with  $y'$  and  $y$  as deviations from  $M_y$  and  $x = X - M_x$ . This is merely the translation of the axes which makes the origin or reference point coincide with  $M_x$  and  $M_y$ . The student should visualize the meaning of this shift of axes. Note that the pattern of tallies is not changed by this simple transformation. Do you think that the slope  $B$  will equal the slope  $b$ ? Will  $A = a$ ? Let us keep the first question in abeyance and examine now the second question. Both  $A$  and  $a$  represent the  $y$  intercepts of the desired prediction line. If it is not immediately obvious to the student that  $A$  may not equal  $a$ , he should imagine that in Table 11 and Fig. 13 the axes have been moved so that the origin is at the center of the scatter diagram, and then ask himself where the line through the means of the vertical arrays would cut the new  $y$  axis. (Incidentally, it should be noted that the value of  $A$  cannot be read directly from Fig. 13 for the simple reason that the reference frame as drawn does not include the origin. The real  $y$  and  $x$  axes of the original measures would be, respectively, to the left of, and lower than, the indicated axes.)

It is of interest to speculate concerning the value of  $a$  in the equation  $y' = bx + a$ . Common sense would suggest that, if an individual were average on  $X$ , the best guess would be that he would be average on  $Y$ . That is, if  $X = M_x$ , one would expect  $Y'$  to equal  $M_y$ . But, if an individual's  $X$  measure fell at  $M_x$ , his deviation, or  $x$  value, would be 0, and the estimated value of  $Y$  as being equal to  $M_y$  would in terms of deviation scores become 0. This would imply that the prediction line would pass through

the origin of the deviation score reference axes, and consequently that the  $y$  intercept would be zero; hence  $a = 0$ . For the purpose of simplifying the determination of the best value for  $b$ , we ask the reader to accept, on the basis of the above reasoning, that  $a = 0$  for the best-fitting line. If we carried both  $a$  and  $b$  along in the following development,  $a$  would in fact turn out to be zero.

This permits us to write  $y' = bx$  as the equation for estimating  $y$ , in deviation units, from  $x$ , or deviation values of  $X$ . Our task becomes that of determining the value of  $b$  which will make  $\Sigma(y - y')^2$  a minimum. Incidentally, it should be obvious that the discrepancy of any particular  $y$  value from the desired line has the same numerical value as the deviation of its corresponding original  $Y$  value from the line, and that  $\Sigma(y - y')^2 = \Sigma(Y - Y')^2$ . When we have determined the optimal value for  $b$  in  $y' = bx$ , we can readily pass back to the original reference frames, the gross score axes, by substituting for  $y'$  the value  $Y' - M_y$ , and for  $x$ ,  $X - M_x$ . With  $a$  fixed as zero, i.e., with the  $y$  intercept equal to zero, we can think of the line as passing through the origin (deviation axes); i.e., its up and down location is fixed. Obviously, many lines could be drawn through the origin, and they would differ only as to slope, i.e., as to  $b$ . Of all possible lines which may be drawn through the origin, some will be closer than others to the observations (tallies) *in toto*. One might imagine several lines any of which would seem to constitute a good fit. As one takes lines with either greater or lesser slope than those of apparently good fit, the fits will become worse; and of those which seem to fit, some will actually be better than others. The student might think that it would only be necessary to draw what seems by inspection to be the best-fitting line, and then obtain its slope by actually measuring the angle which it makes with the horizontal (with needed adjustment to allow for the measurement units). The trouble with this procedure is that individuals would tend to disagree regarding which of several lines was really best; also, the measurement of angles would be none too exact. What we need is a procedure which is objective, a method that will yield the value of  $b$  which leads to the best possible fit in the sense of reducing the sum of the squares of the discrepancies to a minimum.

We set up the function

$$f = \frac{\Sigma(y - y')^2}{N} = \frac{\Sigma(y - bx)^2}{N}$$

in which we have  $N$  deviations of the form  $y - y'$  or  $y - bx$  (since  $y' = bx$ ). These deviations when squared, summed, and divided by  $N$  give us a quantity or function which is to be minimized by the proper choice of  $b$ . The value to be assigned to  $b$  can best be ascertained by the calculus.\* This is done by taking the derivative of the function with respect to  $b$ , setting this derivative equal to zero, and then solving for  $b$ . Thus

$$\frac{df}{db} = \frac{-2\sum x(y - bx)}{N}$$

which, set equal to zero and divided by  $-2$ , gives

$$\frac{\sum x(y - bx)}{N} = 0$$

or

$$\frac{\sum xy - b\sum x^2}{N} = 0$$

then

$$\frac{\sum xy}{N} - b \frac{\sum x^2}{N} = 0$$

The first or cross-product term involves the correlation coefficient as defined by formula (29), from which definition formula we see that  $\sum xy/N = r\sigma_x\sigma_y$ ; and since  $\sum x^2/N = \sigma_x^2$ , we have

$$r\sigma_x\sigma_y - b\sigma_x^2 = 0$$

or

$$r\sigma_y - b\sigma_x = 0$$

which gives

$$b = r \frac{\sigma_y}{\sigma_x}$$

as the optimal value for  $b$ . We therefore have

$$y' = r \frac{\sigma_y}{\sigma_x} x \quad (32)$$

as the equation for the best-fit line. This equation is in terms of

\* The student who has not studied the calculus will either take the first part of the following derivation on faith or, if skeptical, will dig into a calculus text to satisfy himself that no magic is involved here.



1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025, 2026, 2027, 2028, 2029, 2030, 2031, 2032, 2033, 2034, 2035, 2036, 2037, 2038, 2039, 2040, 2041, 2042, 2043, 2044, 2045, 2046, 2047, 2048, 2049, 2050, 2051, 2052, 2053, 2054, 2055, 2056, 2057, 2058, 2059, 2060, 2061, 2062, 2063, 2064, 2065, 2066, 2067, 2068, 2069, 2070, 2071, 2072, 2073, 2074, 2075, 2076, 2077, 2078, 2079, 2080, 2081, 2082, 2083, 2084, 2085, 2086, 2087, 2088, 2089, 2090, 2091, 2092, 2093, 2094, 2095, 2096, 2097, 2098, 2099, 2100, 2101, 2102, 2103, 2104, 2105, 2106, 2107, 2108, 2109, 2110, 2111, 2112, 2113, 2114, 2115, 2116, 2117, 2118, 2119, 2120, 2121, 2122, 2123, 2124, 2125, 2126, 2127, 2128, 2129, 2130, 2131, 2132, 2133, 2134, 2135, 2136, 2137, 2138, 2139, 2140, 2141, 2142, 2143, 2144, 2145, 2146, 2147, 2148, 2149, 2150, 2151, 2152, 2153, 2154, 2155, 2156, 2157, 2158, 2159, 2160, 2161, 2162, 2163, 2164, 2165, 2166, 2167, 2168, 2169, 2170, 2171, 2172, 2173, 2174, 2175, 2176, 2177, 2178, 2179, 2180, 2181, 2182, 2183, 2184, 2185, 2186, 2187, 2188, 2189, 2190, 2191, 2192, 2193, 2194, 2195, 2196, 2197, 2198, 2199, 2200, 2201, 2202, 2203, 2204, 2205, 2206, 2207, 2208, 2209, 2210, 2211, 2212, 2213, 2214, 2215, 2216, 2217, 2218, 2219, 2220, 2221, 2222, 2223, 2224, 2225, 2226, 2227, 2228, 2229, 2230, 2231, 2232, 2233, 2234, 2235, 2236, 2237, 2238, 2239, 2240, 2241, 2242, 2243, 2244, 2245, 2246, 2247, 2248, 2249, 2250, 2251, 2252, 2253, 2254, 2255, 2256, 2257, 2258, 2259, 2260, 2261, 2262, 2263, 2264, 2265, 2266, 2267, 2268, 2269, 2270, 2271, 2272, 2273, 2274, 2275, 2276, 2277, 2278, 2279, 2280, 2281, 2282, 2283, 2284, 2285, 2286, 2287, 2288, 2289, 2290, 2291, 2292, 2293, 2294, 2295, 2296, 2297, 2298, 2299, 2300, 2301, 2302, 2303, 2304, 2305, 2306, 2307, 2308, 2309, 2310, 2311, 2312, 2313, 2314, 2315, 2316, 2317, 2318, 2319, 2320, 2321, 2322, 2323, 2324, 2325, 2326, 2327, 2328, 2329, 2330, 2331, 2332, 2333, 2334, 2335, 2336, 2337, 2338, 2339, 2340, 2341, 2342, 2343, 2344, 2345, 2346, 2347, 2348, 2349, 2350, 2351, 2352, 2353, 2354, 2355, 2356, 2357, 2358, 2359, 2360, 2361, 2362, 2363, 2364, 2365, 2366, 2367, 2368, 2369, 2370, 2371, 2372, 2373, 2374, 2375, 2376, 2377, 2378, 2379, 2380, 2381, 2382, 2383, 2384, 2385, 2386, 2387, 2388, 2389, 2390, 2391, 2392, 2393, 2394, 2395, 2396, 2397, 2398, 2399, 2400, 2401, 2402, 2403, 2404, 2405, 2406, 2407, 2408, 2409, 2410, 2411, 2412, 2413, 2414, 2415, 2416, 2417, 2418, 2419, 2420, 2421, 2422, 2423, 2424, 2425, 2426, 2427, 2428, 2429, 2430, 2431, 2432, 2433, 2434, 2435, 2436, 2437, 2438, 2439, 2440, 2441, 2442, 2443, 2444, 2445, 2446, 2447, 2448, 2449, 2450, 2451, 2452, 2453, 2454, 2455, 2456, 2457, 2458, 2459, 2460, 2461, 2462, 2463, 2464, 2465, 2466, 2467, 2468, 2469, 2470, 2471, 2472, 2473, 2474, 2475, 2476, 2477, 2478, 2479, 2480, 2481, 2482, 2483, 2484, 2485, 2486, 2487, 2488, 2489, 2490, 2491, 2492, 2493, 2494, 2495, 2496, 2497, 2498, 2499, 2500, 2501, 2502, 2503, 2504, 2505, 2506, 2507, 2508, 2509, 2510, 2511, 2512, 2513, 2514, 2515, 2516, 2517, 2518, 2519, 2520, 2521, 2522, 2523, 2524, 2525, 2526, 2527, 2528, 2529, 2530, 2531, 2532, 2533, 2534, 2535, 2536, 2537, 2538, 2539, 2540, 2541, 2542, 2543, 2544, 2545, 2546, 2547, 2548, 2549, 2550, 2551, 2552, 2553, 2554, 2555, 2556, 2557, 2558, 2559, 2560, 2561, 2562, 2563, 2564, 2565, 2566, 2567, 2568, 2569, 2570, 2571, 2572, 2573, 2574, 2575, 2576, 2577, 2578, 2579, 2580, 2581, 2582, 2583, 2584, 2585, 2586, 2587, 2588, 2589, 2590, 2591, 2592, 2593, 2594, 2595, 2596, 2597, 2598, 2599, 2600, 2601, 2602, 2603, 2604, 2605, 2606, 2607, 2608, 2609, 2610, 2611, 2612, 2613, 2614, 2615, 2616, 2617, 2618, 2619, 2620, 2621, 2622, 2623, 2624, 2625, 2626, 2627, 2628, 2629, 2630, 2631, 2632, 2633, 2634, 2635, 2636, 2637, 2638, 2639, 2640, 2641, 2642, 2643, 2644, 2645, 2646, 2647, 2648, 2649, 2650, 2651, 2652, 2653, 2654, 2655, 2656, 2657, 2658, 2659, 2660, 2661, 2662, 2663, 2664, 2665, 2666, 2667, 2668, 2669, 2670, 2671, 2672, 2673, 2674, 2675, 2676, 2677, 2678, 2679, 26

the following is the complete schedule

1" = 241' x 3134' horizontal scale - length

[illegible][illegible]

- The following are some of the most common types of **business** **plans** that you can use to help you decide if you want to start a business. Each plan has its own set of **advantages** and **disadvantages**, so you should carefully consider each one before you decide which one to use. The **business plan** is a document that describes the **business** and its **operations**, and it is used to help you decide if you want to start a business. It is also used to help you decide if you want to raise money from investors or banks. The **business plan** is a document that describes the **business** and its **operations**, and it is used to help you decide if you want to start a business. It is also used to help you decide if you want to raise money from investors or banks.

2. *Journal of Management Education*, 2000, 24(1), 10-12.



we really need something corresponding to the  $\sigma$  about this line. Such a value can be obtained by noting that  $y - y'$  (or  $Y - Y'$ ) represents the discrepancy between estimated and observed values and that  $\Sigma(y - y')^2 / N$  is the mean of the squared deviations, the root of which will be the standard deviation of the discrepancies between estimated and observed values. This will be taken as the one standard deviation to replace the several standard deviations as our measure of the error of prediction. This particular standard deviation, defined as the square root of  $\Sigma(y - y')^2 / N$ , is called the *standard error of estimate*. It may be determined in two ways. First we can take a roundabout way which involves these steps: the prediction of each  $Y$  by use of equation (32a), or each  $y$  by use of (32); the calculation of the discrepancies ( $Y - Y'$ ) or ( $y - y'$ ); squaring, summing, dividing by  $N$ , and taking the square root. A quicker method for determining the standard error of estimate is readily derived algebraically.

Let  $\sigma_{y \cdot x}$  stand for the standard error of  $Y$  as estimated from  $X$ ; then by definition,

$$\sigma_{y \cdot x}^2 = \frac{\Sigma(Y - Y')^2}{N} = \frac{\Sigma(y - y')^2}{N}$$

but

$$y' = r \frac{\sigma_y}{\sigma_x} x$$

by formula (32) whence

$$\begin{aligned} \sigma_{y \cdot x}^2 &= \frac{1}{N} \Sigma \left( y - r \frac{\sigma_y}{\sigma_x} x \right)^2 \\ &= \frac{1}{N} \Sigma \left( y^2 - 2r \frac{\sigma_y}{\sigma_x} xy + r^2 \frac{\sigma_y^2}{\sigma_x^2} x^2 \right) \\ &= \frac{\Sigma y^2}{N} - 2r \frac{\sigma_y}{\sigma_x} \left( \frac{\Sigma xy}{N} \right) + r^2 \frac{\sigma_y^2}{\sigma_x^2} \left( \frac{\Sigma x^2}{N} \right) \\ &= \sigma_y^2 - 2r \frac{\sigma_y}{\sigma_x} r \sigma_x \sigma_y + r^2 \frac{\sigma_y^2}{\sigma_x^2} \sigma_x^2 \\ &= \sigma_y^2 - r^2 \sigma_y^2 \end{aligned}$$

then

$$\sigma_{y \cdot x} = \sigma_y \sqrt{1 - r^2} \quad (34)$$

By a similar line of reasoning it can be shown that

$$\sigma_{x \cdot y} = \sigma_x \sqrt{1 - r^2} \quad (35)$$

which gives the standard error of  $X$  as estimated from  $Y$ .

Thus the correlation coefficient not only enters into the prediction equations (32 to 33a), but also permits us to gauge the accuracy of prediction. It should be noted in passing that one can write the equation of a best-fit line without first determining  $r$  and that the error of prediction can also be ascertained without recourse to  $r$ . Such a method for determining the error of estimate has already been indicated: the square root of  $\Sigma(Y - Y')^2 / N$ , in which  $Y - Y'$  represents the computed discrepancy between observed and predicted values. This need not involve  $r$  unless the prediction equation is written in terms of  $r$ , as was done in (32a). The equation  $Y' = A + BX$  can be written in the form

$$Y' = \frac{\Sigma X^2 \Sigma Y - \Sigma X \Sigma XY}{N \Sigma X^2 - (\Sigma X)^2} + \frac{N \Sigma XY - \Sigma X \Sigma Y}{N \Sigma X^2 - (\Sigma X)^2} (X) \quad (36)$$

in which  $X$  and  $Y$  stand for gross or original measures. Formula (36) for the best-fitting line (least squares solution) does not involve means,  $\sigma$ 's, or the correlation coefficient. If, as is frequently the case, one is interested in obtaining the equation for  $Y$  only, it will be noticed that it is unnecessary to compute the sum of the  $Y$  squares, which is not, however, a tremendous saving of time. Perhaps the quickest way for determining the equation is by direct substitution into (36), but the determination of the error of estimate (sometimes called the closeness of fit of the line) is certainly facilitated by calculating  $r$  and  $\sigma_y$  and substituting in (34).

The standard error of estimate is to be interpreted as a standard deviation, and in so doing we are tacitly assuming that the array distributions are not only equal in dispersion but also normal. For the correlation diagram in Table 11, we have  $\sigma_{y \cdot x} = 1.9$ , which is to be considered the standard deviation of the  $Y$  values about the regression line,  $Y' = .52X + 33.24$ . By use of this equation we would predict that the height of the son of a man 70 inches tall ( $X = 70$ ) would be 69.6, and the error of estimate, 1.9, would be interpreted by saying that, if we made many such predictions, 68.26 times out of a hundred the actual height of sons

## 134 Correlation: Interpretations and Assumptions

of 70-inch fathers would be within the limits  $69.6 \pm 1.9$ , and nearly always within the limits  $69.6 \pm 3(1.9)$ .

This is a *second method* for interpreting the correlation coefficient: in terms of the accuracy of prediction or closeness of fit of regression lines. If no correlation exists, the errors of estimate are  $\sigma_{y \cdot x} = \sigma_y$  and  $\sigma_{x \cdot y} = \sigma_x$ . In this connection it can be seen from formulas (32a) and (33a) that, when  $r = 0$ , the estimated  $Y$ ,  $Y'$ , becomes  $M_y$ , and  $X'$  becomes  $M_x$ . For example, if it has been established that the correlation between toe length and IQ is zero, we would always take 100 (the mean) as our best guess for an individual's IQ regardless of toe length. The error of estimate would of course be the standard deviation of the distribution of IQ's, and it would be said that toe length is useless in predicting IQ. The scatter diagram for IQ as  $Y$  and toe length as  $X$  would exhibit the following characteristics: first, the regression line  $Y' = A + BX$  would be horizontal, i.e.,  $B$  would equal zero, and the means of the arrays would fluctuate about the value  $M_y$ , or  $A$  would equal  $M_y$ ; and, second, all the array distributions would have dispersions approximately equal to  $\sigma_y$ . What would be the best guess as to the other regression line and the standard deviations of the horizontal arrays?

Now suppose the correlation between the variables were perfect ( $r = +1$  or  $-1$ ). The tallies in the scatter diagram would be in a line, there would be no spreading about this line, the two regression lines would coincide, and no error would be involved in estimating  $X$  from  $Y$  or  $Y$  from  $X$ . That  $\sigma_{y \cdot x}$  and  $\sigma_{x \cdot y}$  would both be zero in case of perfect correlation is quite evident when one considers formulas (34) and (35).

At this point the student should note the difference between positive and negative correlation. In the case of a positive  $r$ , a high score goes with high and low with low, whereas, for a negative  $r$ , high goes with low and low with high. With reference to the scatter diagram, a negative  $r$  typically involves a swarm of tallies stretching from the upper-left to the lower-right corner, whereas for a positive  $r$  the trend is from lower left to upper right (this assumes that the axes have been laid off in the conventional fashion). With reference to the regression equations, a negative  $r$  yields negative regression coefficients or negative slope for the lines. The student should be warned that an apparently negative  $r$  may in reality be positive. Thus, if one variable is a test or

performance scored in terms of time (or errors) and the other variable is scored in terms of amount done, the scatter diagram might show large time scores as going with small amounts of work done, i.e., high with low, which might be wrongly taken to indicate negative rather than positive correlation. Instead of asking whether high goes with high and low with low, it is safer to ask whether best goes with best. This rule, however, is difficult to apply when we are dealing with the interrelation of personality traits, especially those which do not readily permit of a statement as to which is the desirable end of the trait scale. The sign of the correlation coefficient in such cases always needs a qualifying statement which explicitly tells the direction of the relationship between the variables. Obviously, as far as accuracy of prediction is concerned, the error is the same for a negative and positive  $r$  of the same magnitude.

**Alienation.** To return to the interpretation of the correlation coefficient by way of the standard error of estimate, we see that the factor in formulas (34) and (35) which involves  $r$  is  $\sqrt{1 - r^2}$ . It is the value of this which, when multiplied by the proper  $\sigma$ , leads to the error of estimate. The expression  $\sqrt{1 - r^2}$  is called the *coefficient of alienation*. If  $r$  is zero, its value is 1 and the error of estimate is the  $\sigma$  for the variable being estimated. Table 12 gives the value of the coefficient of alienation for varying values of  $r$ . The student will do well to fix in mind the trend in this table. It will be noted that, compared to a correlation of zero, an  $r$  of .60 reduces the error of estimate by 20 per cent, whereas an  $r$  of .30 reduces it by about 5 per cent; that  $r$  must be as high as

Table 12. VALUES OF THE COEFFICIENT OF ALIENATION

$r$	$\sqrt{1 - r^2}$	$r$	$\sqrt{1 - r^2}$
.00	1.000	.60	.800
.10	.995	.70	.714
.20	.980	.80	.600
.30	.954	.866	.500
.40	.917	.90	.436
.50	.866	.95	.312

.866 before the error of estimate is reduced by one-half; and that the difference in reduction between an  $r$  of .70 and an  $r$  of .90 is approximately the same as that between .20 and .70. This inter-

pretation of  $r$  is most useful and at the same time most disturbing, since the errors of estimate for  $r$ 's in the vicinity of .40 to .70, values usually found and utilized in predicting success from test results, are discouragingly large.

A somewhat different way of grasping the meaning of  $r$ , as it is applied to accuracy of prediction, is to square both sides of formula (34) and then solve explicitly for  $r$ . This leads to

$$r^2 = 1 - \frac{\sigma_{y \cdot x}^2}{\sigma_y^2}$$

from which it is readily seen that the correlation coefficient depends upon the accuracy of prediction *relative* to the total variance of the variable being predicted.

It might be well at this time to bring together a few remarks concerning the assumptions involved in using and interpreting a correlation coefficient in terms of either rate of change or accuracy of prediction. When an  $r$  is reported, and no evidence to the contrary is given, one has a right to expect that the assumptions of linearity of regression and homoscedasticity have been met. The interpretation of  $r$  as rate of change definitely assumes linearity, and the interpretation in terms of the error of estimate definitely assumes both linearity and homoscedasticity. In certain special cases where the investigator is interested only in a one-way prediction, say  $Y$  from  $X$ , and there is no likelihood of ever reversing to predict  $X$  from  $Y$ , it will suffice if the regression of  $Y$  on  $X$ , i.e., for predicting  $Y$  from  $X$ , be linear and the  $Y$  or vertical array distributions be homoscedastic. The use of the correlation coefficient in predicting performance from age may be cited as an instance in which one need not worry about the possible nonlinear regression of age on score or the lack of homoscedasticity about this regression line.

The student may have observed that no assumptions have been made concerning the nature of the marginal distributions; the utilization of  $r$  does not assume normal distributions for the variables being correlated. The use of the standard error of estimate, however, assumes normality of the array distributions. As regards the possible effect of nonnormal marginal distributions, experience shows that nonlinearity, lack of homoscedasticity, or nonnormality of arrays may frequently be associated with skewness in one or both of the marginal distributions.

Although there are adequate checks for linearity and homoscedasticity, a careful scrutinization of the scatter diagram is usually sufficient to warn one of violent departures from these assumptions. Formula (30) and other nonplotting schemes for computing  $r$  give no inkling as to whether these assumptions are being violated and therefore cannot command the confidence of the careful investigator. The purpose of a research project might very well be the study of the relationship between two variables, but an end result in terms of a correlation coefficient, with no attention given to the form of the relationship, is inadequate.

### VARIANCE AND CORRELATION

A *third method* of interpreting  $r$  is in terms of variance. Before discussing this interpretation, we must introduce an important theorem concerning the variance of a sum (or difference). Suppose that variable  $W$  is made up of two parts  $U$  and  $V$  such that  $W = U + V$ . For example, the score on an arithmetic test might consist of two parts: score in addition and score in multiplication. Obviously,  $w = u + v$ , and therefore the variance of the  $W$  variable is

$$\begin{aligned}\sigma_w^2 &= \frac{\sum w^2}{N} \\ &= \frac{1}{N} \sum (u + v)^2 \\ &= \frac{1}{N} (\sum u^2 + \sum v^2 + 2\sum uv) \\ &= \sigma_u^2 + \sigma_v^2 + 2r_{uv}\sigma_u\sigma_v\end{aligned}\tag{37}$$

and in case  $U$  and  $V$  are independent, we have

$$\sigma_w^2 = \sigma_u^2 + \sigma_v^2\tag{37a}$$

If we are dealing with the difference,  $W = U - V$ , we have

$$\sigma_w^2 = \sigma_u^2 + \sigma_v^2 - 2r_{uv}\sigma_u\sigma_v\tag{38}$$

and for  $U$  and  $V$  independent, we have

$$\sigma_w^2 = \sigma_u^2 + \sigma_v^2$$



which is identical with (37a). In words, *the variance of a sum (or difference) of two independent variables is equal to the sum of their separate variances*. Variances are additive, whereas standard deviations are not. It can be shown that, when  $U$  and  $V$  are distributed normally, their sum or difference will also yield a normal distribution.

Now, with regard to the third method for interpreting  $r$ , let us note that in deviation units an observed  $y$  can be thought of as made up of 2 independent parts, the part which can be predicted from  $x$ , namely  $y'$ , and the residual or unpredictable part,  $(y - y')$ . Before going further we must demonstrate that  $y'$  and  $(y - y')$  are really independent. The numerator for the correlation between  $y'$  and  $(y - y')$  can be expressed as  $\Sigma y'(y - y')$ . But, since  $y' = r \frac{\sigma_y}{\sigma_x} x$  and  $(y - y') = y - r \frac{\sigma_y}{\sigma_x} x$ , we have

$$\begin{aligned}\Sigma y'(y - y') &= \Sigma r \frac{\sigma_y}{\sigma_x} x \left( y - r \frac{\sigma_y}{\sigma_x} x \right) \\ &= r \frac{\sigma_y}{\sigma_x} \Sigma xy - r^2 \frac{\sigma_y^2}{\sigma_x^2} \Sigma x^2 \\ &= r \frac{\sigma_y}{\sigma_x} N r \sigma_x \sigma_y - r^2 \frac{\sigma_y^2}{\sigma_x^2} N \sigma_x^2\end{aligned}$$

which is seen to be zero; hence  $y'$  and  $(y - y')$  are uncorrelated.

We have  $y = y' + (y - y')$ ; whence, by the above variance theorem,

$$\sigma_y^2 = \sigma_{y'}^2 + \sigma_{y \cdot x}^2 \quad (39)$$

in which  $\sigma_{y \cdot x}^2$  is the variance of the residuals,  $(y - y')$ . If we divide both sides of this equation by  $\sigma_y^2$ , we get

$$1 = \frac{\sigma_{y'}^2}{\sigma_y^2} + \frac{\sigma_{y \cdot x}^2}{\sigma_y^2} \quad (39a)$$

from which we see that, since the 2 ratios add to unity, either one can be interpreted as a proportion (or a percentage by shifting the decimal point). Thus the ratio of  $\sigma_{y'}^2$  to  $\sigma_y^2$  is the proportion of the variance in  $Y$  which can be predicted from  $X$ , and the ratio of  $\sigma_{y \cdot x}^2$  to  $\sigma_y^2$  represents the proportion of the variation (variance) of  $Y$  which is left over or remains or cannot be predicted from  $X$ .

A little reflection as to the meaning of this residual variance should convince the student that we are here dealing with the same variance which results if we square formula (34), thus

$$\sigma_{y \cdot x}^2 = \sigma_y^2(1 - r^2)$$

which means that

$$\frac{\sigma_{y \cdot x}^2}{\sigma_y^2} = 1 - r^2$$

When we substitute this value into (39a), we have

$$1 = \frac{\sigma_{y'}^2}{\sigma_y^2} + 1 - r^2$$

from which it is readily seen that the ratio

$$\frac{\sigma_{y'}^2}{\sigma_y^2} = r^2$$

That is, the square of the correlation coefficient gives the proportion of the total variance of  $Y$  which is predictable from  $X$ , or  $r^2$  measures the proportion of the  $Y$  variance which can be attributed to variation in  $X$ . The proportion of the variance of  $Y$  which is due to variables other than  $X$  is given by  $1 - r^2$ . By shifting decimals, we can think of  $r^2$  as indicating a percentage, the percentage of variance which has been explained, and  $1 - r^2$  as the percentage of variance due to other causes. It will be noted that  $r^2$ , not  $r$ , can be so interpreted. This is true because variances are additive, whereas standard deviations are not. It should be emphasized that  $r^2$  as a proportion has to do with variation expressed technically as variance.

It is of some interest to examine the meaning of  $\sigma_{y'}$ . It is the square of the standard deviation of the estimated values, and, with reference to the scatter diagram,  $\sigma_{y'}$  corresponds approximately to what we would obtain if we were to compute the standard deviation about  $M_y$  of the vertical array means, each weighted according to the number of cases in its array. As an exercise, the student can prove  $r^2 = \sigma_{y'}/\sigma_y^2$  by determining directly, rather than by formula (34), that  $\sigma_{y'}^2 = r^2\sigma_y^2$ . (HINT: use the deviation score form of the regression equation.)

This third method of interpreting a correlation coefficient assumes linearity of the regression line involved in predicting  $Y$ , or the dependent variable, from  $X$  as the independent variable;

i.e., the regression of  $Y$  on  $X$  must be linear. If  $X$  were considered as the dependent variable, then the interpretation that  $r^2$  indicates the proportion of the variance of  $X$  explained by  $Y$  would assume linearity for the regression of  $X$  on  $Y$ . The assumption of linearity becomes explicit if one proves directly that  $\sigma^2_{y'} = r^2 \sigma^2_y$ , and it was implied when we used  $\sigma^2_{y \cdot x}$  in that this residual variance was taken about a straight line. This interpretation does not assume homoscedasticity, nor does it assume normality either for the marginal or for the array distributions.

The investigator who is interested in analyzing variation and its possible causes will prefer the interpretation of the correlation coefficient in terms of variance. The problem is frequently one in which an attempt is made to explain variation in one trait in terms of variation of another which is conceived of as being more basic. The use of  $r^2$  as the percentage of the variance of a trait which is predictable by, or attributable to, variation in a second variable becomes a valuable tool in the analysis of variation. Of course one must use caution in assuming causation of one variable by another. Logic, not statistical method, must be invoked to determine whether a causal relationship exists, and the statistical interpretation modified accordingly. Variation in  $X$  might cause variation in  $Y$ , or vice versa, or variation in both  $X$  and  $Y$  might be due to the influence of some other variable or variables.

To illustrate the interpretation of  $r^2$  as a percentage, let us suppose we have the performance of a group of school children on a substitution test. Considerable variation in scores will be present, and we may rightfully ask whether a portion of this variation is due to age differences. We can determine the correlation between age and performance. Suppose  $r = .60$ ; this can be interpreted by saying that 36 per cent of the *variance* in performance is due to age differences, and 64 per cent is due to other causes. Likewise, the variance in crop yield due to variation in rainfall can be determined; or the variance in the height of a group of men may be analyzed into two or more parts, one of which might be the portion due to variation in the heights of their fathers.

### CORRELATION AND COMMON ELEMENTS

A *fourth* possible interpretation of the correlation coefficient assumes that each of the two variables can be thought of as a summation of a number of equally potent, equally likely, inde-

pendent elements, which can be either present or absent. Then the degree of correlation is a function of the number of elements common to the two variables. The general formula is

$$r_{xy} = \frac{n_c}{\sqrt{n_x + n_c} \sqrt{n_y + n_c}} \quad (40)$$

in which  $n_x$  equals the number of elements unique to  $X$ ,  $n_y$  the number unique to  $Y$ , and  $n_c$  the number common to both variables. If the number of elements in  $X$  equals the number in  $Y$ ,  $r$  gives the proportion of elements common to  $X$  and  $Y$ ; if  $X$  is determined only by elements common to  $Y$ , while  $Y$  has additional elements,  $r^2$  gives the proportion of elements entering into  $Y$  which determine  $X$ . There is little, if any, factual basis for believing that the assumptions stated above are tenable so far as psychological variables are concerned, and therefore the interpretation of the correlation coefficient in terms of common elements may be viewed with scepticism.

### NORMAL CORRELATION

A *fifth* interpretation of  $r$  is more mathematical but of little practical value. We have already seen how a frequency distribution and its polygon can be thought of as smooth, conforming perhaps to the equation of the normal curve. A correlation table is a frequency distribution, a picture or graph of which requires a third dimension. If we were to replace each tally in a scatter diagram by a thin block, there would result something analogous to the histogram except that it would be three dimensional—the heights of the stacks of blocks would indicate the frequencies for the various cells. Now suppose that this mound of blocks is by some method smoothed to a surface, and we consider the total volume under the surface (between the surface and the  $XY$  plane) as representing  $N$ . Then the number of cases falling between two given  $X$  values and simultaneously between two given  $Y$  values will be approximately the volume of that portion of the mound which has as its base the rectangle or square formed by the intersections of the two  $X$  and two  $Y$  values. If the regression lines are linear, if the array distributions are normal and homoscedastic, and if the marginal distributions are normal, the resulting surface

is termed the *normal correlation surface*, and the equation of the surface can be written as

$$z = \frac{N}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - \frac{2rxy}{\sigma_x\sigma_y}\right)} \quad (41)$$

A number of important properties of the normal correlation surface can be deduced from this equation and its integral. For instance, the standard error of estimate can be derived from formula (41), and it can also be shown that the contour lines which represent different altitudes on the mound, i.e., different frequencies, will be concentric ellipses, and that if  $r = 0$ , the contour lines will become concentric circles. If the equation is written with  $N$  equal to unity, by double integration the probability of an individual's falling between two particular  $Y$  values and between two  $X$  values can be determined. Tables are available which can be utilized for this purpose.†

### LIMITS FOR $r$

Attention is called to the fact that definition formula (29) becomes  $r = \Sigma z_x z_y / N$ , when written in terms of standard scores for both variables. This indicates specifically that the correlation coefficient is a statistical average, the average of the cross products of standard scores. Suppose that we ask what happens when the correlation is perfect in the sense that each individual's  $z_x$  score equals his  $z_y$  score. If this is true, the sum  $\Sigma z_x z_y$  would be the same as  $\Sigma z^2$ , which when divided by  $N$  gives 1.00. Thus the upper limit for  $r$  is +1.00. Now suppose a perfect inverse relationship, such that an individual's  $z_x$  and  $z_y$  are the same except for sign, one being positive whereas the other is negative. If this holds true for all the cases, the sum  $\Sigma z_x z_y$  can be written as  $\Sigma z(-z)$  or  $-\Sigma z^2$ , which when divided by  $N$  gives -1.00 as the limit for perfect negative correlation.

As exercises, the student should show that multiplying or dividing either  $X$  or  $Y$  or both by a constant, or  $X$  by one constant and  $Y$  by another, will not change  $r$ , and that adding or subtracting a constant does not affect the value of  $r$ .

† Pearson, Karl, *Tables for statisticians and biometricians, part II*, Cambridge: Cambridge University Press, 1931. See Tables 8 and 9.



## SUMMARY

The five suggested methods for interpreting the correlation coefficient may be briefly summarized here.

1.  $r$  is associated with the rate at which one variable changes with another. This assumes that the regression line so interpreted is linear.

2.  $r$  tells us how accurately we can predict by a regression equation. The standard error of estimate permits one to infer the possible magnitude of the prediction error, whereas the coefficient of alienation indicates the reduction in error over that error which would exist if there were no correlation. This interpretation assumes that the regression line used in predicting is linear and that variation about this line is normal and homoscedastic.

3.  $r^2$  gives the proportion of variance in  $Y$  predictable from, or attributable to, variation in  $X$ . This assumes linearity for the regression of  $Y$  on  $X$  and requires caution in assuming the direction of cause and effect.

The student should attempt to visualize the meaning of these three principal methods of interpreting correlation. In particular, he should note the meaning of  $\sigma_y$ ,  $\sigma_{y'}$ , and  $\sigma_{y \cdot x}$  (or their counterparts with the subscripts  $y$  and  $x$  interchanged). The first,  $\sigma_y$ , holds for the marginal distribution of all  $Y$ 's;  $\sigma_{y'}$  pertains to the variability of all  $Y$  values as predicted from  $X$ ; the third,  $\sigma_{y \cdot x}$ , is a measure of the variation about the regression line for predicting  $Y$  from  $X$ .

4.  $r$  or  $r^2$  can be interpreted in terms of the proportion of elements common to the two variables provided we are willing to make rather hazardous and unrealistic assumptions as regards the nature of the variables.

5.  $r$  can be interpreted mathematically in terms of the equation for the normal correlation surface. This assumes that both regressions are linear, that homoscedasticity and normality hold for both the horizontal and vertical array distributions, and that both marginal distributions are normal in form.

The nature of the investigation will usually dictate or suggest the appropriate interpretation. Ordinarily the fifth will not be used in connection with the application of the correlational method, whereas the fourth rests on assumptions which can seldom be met.



## CHAPTER 10

### Factors Which Affect the Correlation Coefficient

Before we interpret, or draw conclusions from, a particular correlation coefficient, it is necessary that we ask ourselves, What factors might have affected its magnitude? The size of an obtained  $r$  depends upon several specific conditions, and, even though it is not always essential that corrections be applied, the investigator must forever be on the lookout for correlations which deviate from their "true" value because of the operation of disturbers. This chapter will be devoted to a discussion of the more common factors which influence  $r$ .

It is assumed that errors in computation have not been permitted—that all arithmetical work has been checked. It is also assumed that sufficient intervals have been used so as to make unnecessary the application of Sheppard's correction for grouping; if more than twelve intervals have been used, the slight increase in  $r$  which results from correcting the standard deviations will be negligible. Certain textbooks have advocated a correction to  $r$  for smallness of the sample, which correction reduces  $r$  by a negligible amount. In view of the magnitude of the effects of other factors on  $r$ , these two possible corrections seem trifling.

#### SELECTION

One of the first questions which must be faced is: Do the cases upon which  $r$  is based represent a random sampling of some defined population, or have selective factors so operated as to increase or decrease  $r$ ? The literature of psychology is not free from correlation coefficients which are decidedly different from values that would have been obtained had the sampling been random. This is not to say that any investigator has willfully selected his

cases so as to produce correlation, but rather to say that unwitting errors are frequently present in spite of an effort to avoid selective factors.

### SAMPLING ERRORS

Even though one feels reasonably sure of the randomness of the sample upon which an  $r$  is based, it is still necessary to consider the obtained  $r$  in terms of variable errors due to sampling. Any  $r$  based on  $N$  pairs of observations will differ more or less from the universe, or population, value,  $\hat{r}$ , which is here conceived of as the value of the correlation coefficient which we would obtain if we had an infinitely large sample. Many of the older texts gave  $(1 - r^2)/\sqrt{N}$  as the standard error of  $r$ , but failed to point out a serious limitation as regards interpretation: that this is an approximation and that  $r$ 's for successive samples are not distributed normally unless  $N$  is large and/or the universe value,  $\hat{r}$ , is near zero.

Before further discussion it should be said that some measure of the sampling fluctuation of the correlation coefficient is highly desirable for any of three reasons: (1) We may wish to say whether an obtained  $r$  can be taken as representing a real, nonchance, correlation, i.e., whether it deviates sufficiently far from zero so that we cannot regard it as a chance fluctuation from no relationship; (2) we may wonder whether a given  $r$  deviates significantly from some a priori or expected value; or (3) we may raise the question of whether two obtained  $r$ 's are significantly different from each other. The answers to these questions must be in terms of probability, and the probability figure which we accept as indicating significance determines the confidence with which we regard any such conclusions as we set forth.

If  $N$  is greater than 30, and if we are interested in saying whether or not an  $r$  (of .50 or less, usually) is significantly different from zero, we can determine its standard error by

$$\sigma_r = \frac{1}{\sqrt{N-1}} \quad (42)$$

and then divide the obtained  $r$  by this standard error in order to secure an  $x/\sigma$  value with which to enter the normal probability table. If  $r/\sigma_r$  is greater than 2.58, we can conclude with a fairly

## 146 Factors Which Affect the Correlation Coefficient

high degree of sureness that the true or universe value of  $r$  is likely to be greater than zero.

For  $N$  less than 30, it is necessary to follow a different procedure. It can be shown that, if the correlation coefficient is computed for successive samples drawn from a population for which the correlation is zero, the successive values of

$$t = r \frac{\sqrt{N-2}}{\sqrt{1-r^2}} = \frac{r}{\sqrt{\frac{1-r^2}{N-2}}}$$

will follow the  $t$  distribution with  $df = N - 2$ . If a sample  $t$  reaches the .01 level of significance, one would conclude that it is not a chance deviation from zero, or that some correlation exists between the 2 variables involved.

From the foregoing expression, it would appear that the  $t$  for testing the significance of correlation is nothing more than an  $r/s_r$ , with  $s_r = \sqrt{(1-r^2)/(N-2)}$  as an estimate of the sampling error of  $r$ . However, there are subtle mathematical reasons why such an interpretation is not permissible.

The student may wonder why the  $df$  is taken as  $N - 2$ . Actually, when we test the significance of an  $r$ , we are testing the significance of regression. If  $r$  is zero, the regression is zero in the sense that the regression coefficient or slope of the regression line is zero. Now a linear regression line involves 2 constants, its slope and its intercept; hence 2 degrees of freedom are lost in fitting the line. Suppose  $N = 2$ , and that the 2  $X$  scores differ; likewise, the 2  $Y$  scores. Imagine these pairs of scores plotted in a scatter diagram, and a regression line fitted or a correlation coefficient computed. The regression line would go through both plotted points; therefore for the sample of 2 cases the prediction would be perfect and  $r$  would be unity. The student may, as an exercise, prove algebraically that, when  $N = 2$  and when there is variation in both  $X$  and  $Y$ , the correlation must be  $+1$  or  $-1$ . In other words, with  $N = 2$  there is no freedom for sampling variation in the numerical value of  $r$ .

Formulas for the standard error of  $r$ , when  $\hat{r}$  is large, are misleading because for high values of  $\hat{r}$  the distribution of successive sample values is markedly skewed. This skewness becomes noticeable when  $\hat{r}$  reaches .40 or .50 and increases rapidly as  $\hat{r}$

nears unity. The skewness is also a function of  $N$ . Because of this skewness the standard error of  $r$  loses its meaning; it cannot be expected to yield a trustworthy answer as to whether an obtained  $r$  deviates significantly from some a priori value, nor can the significance of the difference between 2  $r$ 's be determined by substituting in the ordinary formula for the standard error of a difference.

**The  $r$  to  $z$  transformation.** Professor R. A. Fisher has developed a very useful and accurate technique for handling sampling errors for high values of  $r$ . This procedure is also applicable for low  $r$ 's and can be used when  $N$  is large or small. He employs a transformation

$$z = \frac{1}{2} \log_e (1 + r) - \frac{1}{2} \log_e (1 - r) \quad (43)$$

or

$$z = 1.1513 \log_{10} \frac{1 + r}{1 - r} \quad (43a)$$

which has 2 distinct advantages: (1) the distribution of  $z$  for successive samples is independent of the universe value,  $\hat{r}$ ; i.e., for a given  $N$  the sampling distribution will have the same dispersion for all values of  $\hat{r}$ ; (2) the distribution of  $z$  for successive samples is so nearly normal that it can be treated as such with very little loss of accuracy. The standard error of  $z$  is

$$\sigma_z = \frac{1}{\sqrt{N - 3}} \quad (44)$$

If we wish to state the .99 confidence limits for  $\hat{r}$ , we transform the obtained  $r$  to  $z$  by formula (43a) or by Table B of the Appendix, determine  $\sigma_z$ , find  $z + 2.58\sigma_z$  and  $z - 2.58\sigma_z$ , and then transform these 2  $z$  values back to  $r$ 's by using Table C. As an example and in contrast to the less exact procedure of taking  $r \pm 2.58\sigma_r$ , where  $\sigma_r = (1 - r^2)/\sqrt{N}$ , let us suppose an  $r$  of .90 based on an  $N$  of 50. The standard error of  $r$  by the usual formula is .027; whence  $.90 \pm (2.58)(.027)$  yields the values .830 and .970 as confidence limits for the universe value. Now, if we utilize the  $z$  transformation, we find  $z = 1.47$ , and  $\sigma_z = .146$ , whence  $1.47 \pm (2.58)(.146)$  gives 1.093 and 1.847. These 2 values are then transformed back to the 2  $r$  values, .798 and .951, which it will be noted differ from the confidence limits for  $\hat{r}$  as determined by the classical method.

## 148 Factors Which Affect the Correlation Coefficient

**Difference between  $r$ 's.** If we wish to determine the significance of the difference between 2  $r$ 's, both are transformed into  $z$ 's, and the standard error of the difference between the 2  $z$ 's is obtained by

$$\sigma_{z_1 - z_2} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}} \quad (15)$$

and then the ratio of the difference to its standard error is treated in the usual manner. If the  $z$ 's are significantly different, we conclude that the 2  $r$ 's are significantly different.

Suppose we have the correlation between  $X_1$  and  $X_2$  and also between  $X_1$  and  $X_3$ , with both  $r$ 's based on the same sample of  $N$  cases, and we wish to decide whether there is a significant difference between  $r_{12}$  and  $r_{13}$ . The foregoing method is not applicable because we need to allow for the fact that, for successive samplings,  $r_{12}$  and  $r_{13}$  are not independently distributed, but correlated. The standard error of the difference must include a subtractive  $r$  term involving the correlation between the correlation coefficients. The methods for estimating this needed correlation are none too satisfactory, but there is a test which is interpretable by way of the  $t$  table for  $N$  small and by way of the normal table for  $N$  large. It has been shown that

$$t = \frac{(r_{12} - r_{13})\sqrt{(N-3)(1+r_{23})}}{\sqrt{2(1-r_{12}^2-r_{13}^2-r_{23}^2+2r_{12}r_{13}r_{23})}}$$

follows the  $t$  distribution with  $N-3$  degrees of freedom when the null hypothesis of no difference is true. If  $t$  is significant we conclude that one variable correlates higher than the other with  $X_1$ .

**Averaging correlations.** When we have 2 or more sample values for the correlation between 2 variables we may wish to average the  $r$ 's. (1) in case it is known that the samples have been drawn from the same population or (2) in case it can be assumed because the  $r$ 's are not significantly different from each other that the samples have been drawn from equally correlated populations. An appropriate procedure is to convert each  $r$  to  $z$ , then take a weighted  $z$  by the inverse of its sampling variance (average of the  $z$ 's). Thus, for 3 sample values this weighted average is given by

$$z_{av} = \frac{(N_1-3)z_1 + (N_2-3)z_2 + (N_3-3)z_3}{(N_1-3) + (N_2-3) + (N_3-3)}$$

Then  $r_{xy}$  can be transformed back to an  $r$ , and any significance test concerning such an average  $r$  would be made on  $\Sigma r$ , which has a standard error of

$$1/\sqrt{(N_1 - 3) + (N_2 - 3) + (N_3 - 3)}$$

### RANGE OR SPREAD OF TALENT

The magnitude of the correlation coefficient varies with the degree of heterogeneity with respect to the traits being correlated of the sample. If we are drawing a sample from a group which is restricted in range with regard to either or both variables, the correlation will be relatively low. Thus the restricted range of intelligence is one factor which leads to lower correlation between intelligence and grades for college students than that normally found for high school groups. If the range with respect to 1 variable has been curtailed, and one knows the standard deviation for an uncurtailed distribution, it is possible to adjust the correlation for the difference in range, provided one can be sure of the tenability of 2 assumptions: that the regressions are linear, and that the arrays are homoscedastic for the scatter based on the uncurtailed distribution. If the curtailment is in variable  $X$ , and we let

$\sigma_x$  = SD for curtailed distribution,

$\Sigma_x$  = SD for uncurtailed distribution,

$r_{xy}$  = correlation of variable  $Y$  with  $X$  for curtailed range,

$R_{xy}$  = correlation of variable  $Y$  with  $X$  for uncurtailed range,

the relationship by which we would predict  $R_{xy}$  from  $\sigma_x$ ,  $\Sigma_x$ , and  $r_{xy}$  is given by

$$R_{xy} = \frac{r_{xy}(\Sigma_x/\sigma_x)}{\sqrt{1 - r_{xy}^2 + r_{xy}^2(\Sigma_x/\sigma_x)^2}} \quad (46)$$

Obviously, if we have  $R$  instead of  $r$ , the value of  $r$  for a restricted range can be estimated by formula (46). All we need to do is interchange  $\Sigma$  and  $\sigma$ , and  $r$  and  $R$ , and  $r$  can then substitute for  $R$ . The estimation of  $r$  need not be made in ignorance of whether the assumptions of linearity and homoscedasticity can be met; an examination of the associated scatter for the uncurtailed range will reveal the facts.



## 150 Factors Which Affect the Correlation Coefficient

Formula (46) indicates definitely that the magnitude of the correlation coefficient is a function of the degree of heterogeneity with respect to one of the traits being correlated. A better appreciation of the extent of this influence can be had by examining Table 13 which gives, for varying values of  $R_{xy}$  along the top and

Table 13. VALUES FOR  $r_{xy}$  FOR  $R_{xy}$ 'S OF .30, .40, ... .80 WITH  $\sigma_x/\Sigma_x$  VALUES OF .90, .80, ... .50

$\sigma_x/\Sigma_x$	$R_{xy}$					
	.30	.40	.50	.60	.70	.80
.90	.272	.366	.461	.559	.662	.768
.80	.244	.330	.419	.514	.617	.730
.70	.215	.292	.375	.465	.566	.682
.60	.185	.253	.327	.410	.507	.625
.50	.155	.213	.277	.351	.440	.555

different  $\sigma_x/\Sigma_x$  ratios along the left, the corresponding values of  $r_{xy}$ . It can be shown that double selection, i.e., curtailment on both variables, tends to depress the correlation coefficient. Since the formulas for "correcting" for double curtailment are not too satisfactory, none is given here.

One important rule emerges from the foregoing: standard deviations should always be reported along with correlation coefficients, and some indication should be given as to variation typically found for the variables.

### EFFECT OF UNRELIABILITY

Before considering the effect of unreliability, or errors of measurement, upon the correlation between 2 variables, it is necessary that we digress to explain briefly what is meant by reliability. If we were assigned the task of determining the height of an individual by the use of a tape measure, we might be satisfied with 1 measurement, but unfortunately a single determination might not be entirely free from error. To overcome this, 2 or more measures are averaged on the assumption that the chance or *variable* errors will more or less cancel out. If one computes the standard deviation of the distribution of several measurements (of the same thing), a summary figure indicating the possible magnitude of the variable errors will be obtained. This  $\sigma$  neither

pertains to nor measures the magnitude of a possible *constant* error, i.e., an error which affects all the measurements in the same direction. We are here concerned only with the magnitude of variable errors, or inaccuracies in measurement which are of a chance nature.

**Reliability.** If we had the problem of determining the error in the measurement of height, we could make several measurements on 1 person and compute a measure of accuracy, or we might make just 2 measures on each of several persons and take some function of the difference between the 2 measurements for all  $N$  individuals as our gauge of accuracy. Either scheme leads to an estimate of the size of the variable errors that may be involved.

In psychological measurement, it is not always feasible or possible to obtain more than 2 measures on an individual for a given trait; hence it is necessary to use the second-mentioned scheme for determining the accuracy of measurement. The mean or median *absolute* error may suffice, but, as in physical measurement, we sometimes need to know the extent of the variable errors in relation to the magnitude of the thing being measured, i.e., the *relative* or percentage error. Psychologists have found it useful to interpret variable errors, not with regard to the magnitude (a nearly meaningless word in psychological tests) of the measures, but relative to the variability of the trait for a specific group of individuals. The correlation between 2 determinations is, as we shall soon see, one method of expressing the accuracy of measurement relative to the trait dispersion. Such a correlation is termed the *reliability coefficient*.

Suppose  $X$  = an obtained score or measure for an individual.

$X_{\infty}$  = his true score.

$e$  = a variable error, positive or negative.

Then we can consider that

$$X = X_{\infty} + e$$

or in deviation units

$$x = x_{\infty} + e$$

The variance of the obtained scores will be

$$\sigma_x^2 = \sigma_{\infty}^2 + \sigma_e^2 \quad (47)$$

provided we can assume  $x_{\infty}$  and  $e$  uncorrelated. This assump-

## 152 Factors Which Affect the Correlation Coefficient

tion seems reasonable since the variable error,  $e$ , is supposed to be a chance affair, as often positive as negative, and therefore its magnitude and direction should not be related to anything else. Equation (47) can be stated in words: the variance of the distribution of scores can be broken up into 2 portions, the variance of the true scores and the variance due to errors of measurement.

Suppose that for a given trait we have 2 measurements, each of which is in error but not necessarily to the same extent or in the same direction. Symbolically,

$$x_1 = x_{\infty} + e_1$$

$$x_2 = x_{\infty} + e_2$$

in which the  $e$ 's represent the errors which go with the 2 obtained scores. The reliability coefficient is defined as the correlation between 2 comparable measures of the same thing, i.e., the correlation between  $x_1$  and  $x_2$ . (Each measured individual will have an  $x_1$  and an  $x_2$  score.) Thus we have the reliability coefficient,

$$\begin{aligned} r_{11} = r_{x_1 x_2} &= \frac{\Sigma x_1 x_2}{N \sigma_1 \sigma_2} = \frac{\Sigma (x_{\infty} + e_1)(x_{\infty} + e_2)}{N \sigma_1 \sigma_2} \\ &= \frac{\Sigma x_{\infty}^2 + \Sigma x_{\infty} e_2 + \Sigma x_{\infty} e_1 + \Sigma e_1 e_2}{N \sigma_1 \sigma_2} \end{aligned}$$

Dividing by  $N$  gives

$$r_{11} = \frac{\sigma_{x_{\infty}}^2 + r_{x_{\infty} e_2} \sigma_{x_{\infty}} \sigma_{e_2} + r_{x_{\infty} e_1} \sigma_{x_{\infty}} \sigma_{e_1} + r_{e_1 e_2} \sigma_{e_1} \sigma_{e_2}}{\sigma_1 \sigma_2} \quad (48)$$

If we assume all three  $r$ 's in the numerator equal to zero, we have

$$r_{11} = \frac{\sigma_{x_{\infty}}^2}{\sigma_1 \sigma_2}$$

It is assumed that we are correlating *comparable* measures of the *same* thing or trait—comparable in the sense that  $\sigma_{e_1} = \sigma_{e_2}$ , and  $\sigma_1 = \sigma_2$ . (The *same* trait is implied in that  $x_1$  and  $x_2$  are measures of  $x_{\infty}$ .) Whence we have

$$r_{11} = \frac{\sigma_{x_{\infty}}^2}{\sigma_x^2} \quad (49)$$

where  $\sigma_x = \sigma_1 = \sigma_2$ . The reliability coefficient can be interpreted as a proportion, since from formula (47) we have

$$\frac{\sigma_\infty^2}{\sigma_x^2} + \frac{\sigma_e^2}{\sigma_x^2} = 1$$

i.e., the reliability coefficient represents the proportion of the variance of the obtained scores which is due to the variance of the true scores. It follows that  $1 - r_{11}$  gives the proportion of the variance which is due to errors of measurement.

Obviously, the reliability coefficient can, by substitution from formula (49) into the above expression, also be written as

$$r_{11} = 1 - \frac{\sigma_e^2}{\sigma_x^2} \quad (50)$$

which indicates clearly that the reliability coefficient is a function of the magnitude of the variable error *relative* to the variability of the trait in question. It also follows from formula (50) that the error of measurement can be stated in terms of the reliability coefficient and  $\sigma_x$ ; thus,

$$\sigma_e = \sigma_x \sqrt{1 - r_{11}} \quad (51)$$

That  $\sigma_e$  is to be interpreted as the *standard error of measurement* may be clarified if we note that, when  $x$  ( $= x_1$  or  $x_2$ ) is taken as evidence of the true score,  $x - x_\infty$  becomes the error, and the standard deviation of such errors will be  $\sigma_e$ , as can be shown by easy algebra (an exercise). If it were possible to secure a large number of measures on an individual, we would expect these measures to distribute themselves normally about the true score with a standard deviation corresponding to  $\sigma_e$ . Thus, if the result of one testing yields an IQ of 80, and if  $\sigma_e = 3$ , we can conclude with high confidence that an individual's true position, on the scale of measured (obtained) IQ's, is somewhere between 71 and 89 ( $80 \pm 3\sigma_e$ ), and with fair confidence that it is somewhere between 74 and 86. It can readily be seen that the error of measurement expressed as a  $\sigma$  has a distinct advantage over such concepts as the mean or median error, or the mean or median difference between 2 measurements, in that  $\sigma_e$  enables us to use the probability table either in establishing confidence limits or in deter-

## 154 Factors Which Affect the Correlation Coefficient

mining whether 2 scores differ more than is to be expected on the basis of chance.

There are 2 distinct situations for which one may wish to say whether 2 scores differ more than expected on the basis of error. First, consider 2 individuals each with a score on a given test. The standard error of the difference between the scores is given by  $\sqrt{\sigma_e^2 + \sigma_e^2} = \sigma_e \sqrt{2}$ . Second, consider 1 person with some type of comparable standard scores,  $Z_x$  and  $Z_y$ , on 2 tests having reliability coefficients,  $r_{xx}$  and  $r_{yy}$ , and errors represented by  $\sigma_{e_x}$  and  $\sigma_{e_y}$ . The standard error (of measurement) for the difference score,  $D = Z_x - Z_y$ , is  $\sqrt{\sigma_{e_x}^2 + \sigma_{e_y}^2} = \sigma \sqrt{1 - r_{xx} + 1 - r_{yy}}$  where  $\sigma$  is the standard deviation common to the 2 sets of standard scores. For either situation, a difference between obtained scores divided by the appropriate standard error of the difference provides a *Cr* for judging significance.

Since difference scores of the second type frequently enter into clinical diagnosis or are used as a basis for guidance, it is of interest to know that the reliability coefficient for difference scores is

$$r_{dd} = \frac{r_{xx} + r_{yy} - r_{xy}}{2 - 2r_{xy}} \quad (52)$$

Even though both  $r_{xx}$  and  $r_{yy}$  are satisfactorily high, the difference scores may have far from satisfactory reliability. If, for example,  $r_{xx} = r_{yy} = .90$ , and  $r_{xy} = .70$ , the value of  $r_{dd}$  is only .67.

**Determination of reliability.** The above argument regarding the interpretation of the reliability coefficient either as an indicator of relative accuracy or in terms of  $\sigma_e$  rests on the supposition that we have obtained the reliability coefficient as the result of correlating *comparable* measures of the *same* thing and that the *variable errors are uncorrelated* with themselves and with the true scores. The practical determination of the reliability coefficient involves more, therefore, than the mere correlating of 2 sets of measurements. The conditions under which the 2 sets of scores are obtained must be scrutinized for possible violation of the requisite assumptions. Some of the difficulties involved in ascertaining the reliability of a psychological measurement are suggested in the following paragraphs.

First let us note that the chance variable error,  $e$ , can be broken up into many smaller components at least logically, although not

necessarily experimentally. Thus we might set

$$e = e_a + e_b + e_c + e_d + e_f + \dots$$

in which  $e_a$  = error in the instrument or test

$e_b$  = error due to extraneous physical disturbance

$e_c$  = error due to physiological condition of individual.

$e_d$  = error in scoring or in reading instrument

$e_f$  = error due to day-to-day fluctuations

Other sources of variable error might be added, or some of those listed might be broken up into more minute parts. It is not assumed that these several sources contribute an equal amount to the variance of  $e$ , nor is it assumed that these several components are entirely independent of each other. For instance, daily fluctuations might be influenced by physiological condition.

The assumption of uncorrelated errors implies that  $e_1$  is not correlated with  $e_2$ . Of course the 2 scores for an individual might by chance contain a variable error of the same magnitude and sign, we are here interested, however, in whether an error which is chance for one score might tend in general to affect the second score in the same manner. For example, an upset stomach might lead to a reduced performance score, and if the second test was administered the same day, this same chance factor would affect the second performance score in the same direction. Thus in examining any proposed scheme for determining the reliability of a test we must inquire as to whether any of the sources of error can affect the 2 measurements on an individual in the same direction. If it seems reasonable to suspect that errors are correlated, it follows that the obtained reliability coefficient will be spuriously high since the presence of correlated errors will not allow formula (18) to be reduced to (19).

Let us consider a few of the "accepted" schemes for ascertaining reliability in order to see whether they are "acceptable" in light of the assumptions requisite to a sound reliability coefficient. These assumptions may be recapitulated in the form of 3 questions. Do the 2 tests or determinations represent measures of the same thing? Are the 2 series of measures comparable, comparable tests or instruments? Is it possible or likely that the errors of measurement are correlated, i.e., can the error on the



first test be correlated with the error on the second, or can the error on either be correlated with the true measure?

For the ordinary mental, personality, or achievement test, reliability is usually ascertained by correlating supposedly equivalent (comparable?) forms, by correlating split halves (odd vs. even items or first half vs. second half of test), or by correlating test-retest scores. The test-retest method is of limited value in that there may be a memory carry-over from test to retest, in which case the retest will measure the same trait as the original test plus memory effects. In order to overcome this memory transfer, the retest may be administered some months after the first test, but this permits of a possible change in the trait or ability as a result of maturation or experience.

Split-half reliability involves the correlating of 2 halves and applying the Brown-Spearman formula to determine the reliability of scores based on the whole test. This formula is easily derived. Let  $X_1$  and  $X_2$  stand for the respective halves. Now  $r_{12}$  would be the reliability for scores based on either half, but in practice we always use total scores, defined as  $X_a = X_1 + X_2$ . The reliability of  $X_a$  can be thought of in terms of the correlation between  $X_a$  and an imaginary set of comparable scores,  $X_b = X_3 + X_4$ , where  $X_3$  and  $X_4$  are scores on the 2 respective halves of a nonexistent form of the test. Given information about  $X_1$  and  $X_2$ , we seek an expression for  $r_{ab}$ . In deviation units,  $x_a = x_1 + x_2$  and  $x_b = x_3 + x_4$ ; hence we may write

$$\begin{aligned} r_{ab} &= \frac{\Sigma x_a x_b}{N \sigma_a \sigma_b} = \frac{\Sigma (x_1 + x_2)(x_3 + x_4)}{N \sigma_a \sigma_b} \\ &= \frac{\Sigma x_1 x_3 + \Sigma x_1 x_4 + \Sigma x_2 x_3 + \Sigma x_2 x_4}{N \sigma_a \sigma_b} \end{aligned}$$

Dividing through by  $N$  and utilizing formula (29), and with formula (37) as a basis for specifying  $\sigma_a$  and  $\sigma_b$ , we have

$$r_{ab} = \frac{r_{13}\sigma_1\sigma_3 + r_{14}\sigma_1\sigma_4 + r_{23}\sigma_2\sigma_3 + r_{24}\sigma_2\sigma_4}{\sqrt{\sigma_1^2 + \sigma_2^2 + 2r_{12}\sigma_1\sigma_2} \sqrt{\sigma_3^2 + \sigma_4^2 + 2r_{34}\sigma_3\sigma_4}}$$

Now it is assumed that the  $X_1$  and  $X_2$  scores are comparable (equivalent sets, with  $\sigma_1 = \sigma_2$ ), and we simply say that our imaginary scores,  $X_3$  and  $X_4$ , are comparable with each other and also

with  $X_1$  and  $X_2$ ; hence all 4  $\sigma$ 's have the same value, and therefore cancel out, leaving

$$r_{ab} = \frac{r_{13} + r_{14} + r_{23} + r_{24}}{\sqrt{2 + 2r_{12}}\sqrt{2 + 2r_{34}}}$$

Comparable or equivalent sets of scores will correlate equally with each other, that is, the 5 unknown  $r$ 's in this expression will all equal  $r_{12}$ , our known value. Therefore we have

$$r_{ab} = \frac{4r_{12}}{\sqrt{2 + 2r_{12}}\sqrt{2 + 2r_{12}}} = \frac{2r_{12}}{1 + r_{12}} \quad (53)$$

as the reliability of scores based on the whole test.

The only assumption underlying formula (53) is that the 2 halves being correlated are comparable (equivalent or parallel). If the test items have been arranged according to difficulty, a first-half vs. second-half reliability will not satisfy the notion of comparable measures. Ordinarily the odd-even item technique will satisfy the criteria of comparability and sameness of trait. Neither of the split-half methods will satisfy the assumption of uncorrelated errors. Since both measures are determined at the same sitting, any chance fluctuations due to physiological conditions or to chance factors in the test situation will influence the 2 scores of an individual in the same direction. It is to be expected, therefore, that the correlation of halves will in general lead to a reliability coefficient which is too high, giving us an exaggerated notion of the accuracy with which we can place an individual on the trait continuum.

By far the best method for determining the reliability of a test is to have 2 forms which have been made equivalent and comparable by careful selection and balancing of items. No item in one form should be so nearly identical with an item in the other form as to permit a direct memory transfer. Two forms, equivalent yet not identical, can be administered within, say, 2 weeks' time—a procedure which properly includes in the estimate of variable error the daily fluctuations due to either physiological or psychological conditions and variations due to chance factors in the physical situation in which the tests are given. With so short an interval between testings, the trait being measured will

## 158 Factors Which Affect the Correlation Coefficient

have changed only a negligible amount as a result of maturation or ordinary environmental influences.

When we attempt to obtain the reliability of a learning score or of any performance which is influenced by practice, we encounter difficulties which are baffling to the researcher who rigorously adheres to the fundamental requisites of the reliability coefficient. The chief difficulty is the obvious fact that the "thing" being measured changes as a result of each measurement or trial. Test-retest, or first half vs. second half (of trials), or today's trials vs. tomorrow's will not represent measures of the same function, nor will any scheme analogous to equivalent forms avoid this difficulty, since "forms" which are comparable will permit transfer. The use of scores on odd vs. even trials will have the advantage of balancing somewhat the influence of practice, especially if several trials are given; but the possibility that a chance error affects odds and evens alike is present, in that a slip in the experimental procedure or a temporary discouragement on the part of the testee or the adoption by the subject of a poor approach to the problem will have a similar effect on both scores. If trials were spaced, say, a day apart, the factors just mentioned might not greatly disturb the reliability determination. In general, it can be said that the odd-even trial method will yield a reliability coefficient which is higher than the "true" reliability.

The same shortcomings are present in the aforementioned methods when they are employed in determining the reliability of animal (or human) maze-learning scores. Other techniques, peculiar to the maze situation, have been proposed. Performances on the odd and even blinds, somewhat similar to odd and even items, have been correlated for the purpose of reliability, but since blinds differ considerably as regards difficulty, one cannot be sure that the 2 halves are comparable. One can also question the comparability of the first half and second half of the maze, since in general the last part tends to be learned more quickly than the first. Attempts to ascertain the reliability of one maze by correlating performance on it with that on another maze involve several difficulties. In the first place, there seems to be a general positive transfer (perhaps a general adaptation to the maze situation) from a first to a second maze; secondly, the second maze must be similar to the first in order to satisfy the requisite of comparable measures of the same ability, but if this similarity

approaches identity the second maze becomes a retest; and thirdly, a close degree of similarity will lead to possible interference effects which may act differentially from animal to animal.

The foregoing brief discussion of the requisites for, and difficulties in arriving at, a meaningful reliability coefficient should make obvious the necessity for examining critically any proposed method of determining the reliability of a psychological measurement. The interpretation of the reliability coefficient in terms of the standard error of measurement definitely assumes homoscedasticity, which is another way of saying that the reliability coefficient is valid only when the error of measurement is of the same order of magnitude for the entire range of scores. That this may not always hold true is evident from findings with the 1937 Stanford Revision of the Binet Test.

It should be noted in passing that the magnitude of the reliability coefficient is influenced by the trait homogeneity of the sample upon which it is based. Let  $\sigma$  represent the standard deviation for the restricted range,  $\Sigma$  the standard deviation for the unrestricted range,  $r_{11}$  the reliability for the restricted, and  $R_{11}$  the reliability for the unrestricted; it may be assumed that  $\sigma_e$  for the smaller range equals  $\sigma_e$  for the larger range, i.e.,

$$\sigma^2(1 - r_{11}) = \Sigma^2(1 - R_{11}) \quad (54)$$

This is the usual formula given for relating reliability to amount of variability. It can be argued, however, that the important consideration is the relationship of the reliability coefficient to the amount of true variance; hence the formula should be written in the form

$$\sigma^2 r_{11}(1 - r_{11}) = \Sigma^2 R_{11}(1 - R_{11}) \quad (54a)$$

**Attenuation.** Now we return to the question which led to this lengthy detour: How does unreliability affect the correlation between variables? Let

$$x = x_{\infty} + e$$

$$y = y_{\infty} + d$$

where  $e$  and  $d$  represent the variable errors in the two scores,  $x$  and  $y$ . Then

$$r_{xy} = \frac{\Sigma(x_{\infty} + e)(y_{\infty} + d)}{N\sigma_x\sigma_y}$$

$$= \frac{\Sigma x_{\infty}y_{\infty} + \Sigma x_{\infty}d + \Sigma y_{\infty}e + \Sigma ed}{N\sigma_x\sigma_y}$$

If we assume that  $d$  is uncorrelated with  $x_{\infty}$ , that  $e$  is uncorrelated with  $y_{\infty}$ , and that  $e$  and  $d$  are uncorrelated, we have

$$r_{xy} = \frac{\Sigma x_{\infty}y_{\infty}}{N\sigma_x\sigma_y} = \frac{r_{x_{\infty}y_{\infty}}\sigma_{x_{\infty}}\sigma_{y_{\infty}}}{\sigma_x\sigma_y} \quad (r_{x_{\infty}y_{\infty}} = r \text{ between true scores})$$

Since  $\sigma_{\infty} = \sigma\sqrt{r_{11}}$  by formula (49),

$$r_{xy} = \frac{r_{\infty\infty}\sigma_x\sqrt{r_{x_1x_2}}\sigma_y\sqrt{r_{y_1y_2}}}{\sigma_x\sigma_y}$$

$$= r_{\infty\infty}\sqrt{r_{x_1x_2}}\sqrt{r_{y_1y_2}} \quad (55)$$

which, since the reliability coefficients are less than unity, shows clearly that the correlation between obtained scores will be less than that between true scores; i.e., errors of measurement tend to reduce or attenuate the correlation between traits.

One can rearrange formula (55) as

$$r_{\infty\infty} = \frac{r_{xy}}{\sqrt{r_{x_1x_2}}\sqrt{r_{y_1y_2}}} \quad (56)$$

by which one can estimate what the correlation would be if perfect, errorless, measures were available. This is known as *correction for attenuation*. Correlation coefficients corrected for attenuation are of theoretical importance in the analysis of relationships in that allowance can be made for variable errors of measurement, but such corrected  $r$ 's are of little practical value since they cannot be used in prediction equations. The prediction of one variable from another and the accompanying error of estimate must necessarily be based on obtained, or fallible, rather than true scores.

Since the correlation between variables is a function of the reliability of their measurement, we may examine the limits imposed upon  $r$  as a result of fallible scores. By reference to formula (55), we observe that, if the correlation between true scores is unity and if the reliability for 1 variable is perfect, the obtained

correlation between the 2 cannot exceed the square root of the reliability coefficient for the other variable. If the correlation between the true scores is perfect and if each variable is subject to errors of measurement, then the obtained correlation cannot exceed the product of the square roots of the 2 reliability coefficients. Obviously, if the reliabilities are the same, the obtained correlation cannot be greater than the reliability coefficient.

In addition to the assumptions which were made specifically in deriving the formula for correcting for attenuation, it is also necessary to meet all the assumptions required for a sound reliability coefficient. Since obtained correlations and also reliability coefficients are functions of the homogeneity, with respect to the 2 traits, of the sample upon which they are based, it follows that the reliability coefficients used in correcting an obtained  $r$  should be based on the same sample as  $r$  or on a sample which is of comparable homogeneity. Corrected  $r$ 's greatly in excess of unity have been reported. Such absurd results lead one to ask whether the assumptions have been met, but this question should be raised concerning any corrected  $r$ , even though it does not exceed unity, since the assumptions are difficult to meet. It has been said that a corrected  $r$  can legitimately exceed unity by as much as 2 or 3 times its sampling error. Formulas for the standard error of a corrected  $r$  are available, but nothing is known concerning the nature of the distribution of corrected  $r$ 's for successive samples. Presumably this distribution would be markedly skewed for high values; hence the use of an ordinary standard error technique to determine whether a corrected  $r$  exceeds unity (or any other magnitude) by more than can reasonably be expected on the basis of sampling is an unsound procedure.

## INDEX CORRELATION

A possible source of error in correlational work may be introduced when 2 indexes having a common variable denominator are correlated, such as  $X/Z$  and  $Y/Z$ . Before considering this special case, it might be well to turn our attention to more general formulas for indexes. These formulas involve the coefficient of variation, namely,  $v = \sigma/M$ , and their use leads to serious error when the  $v$ 's are large— $v^3$  and higher-power terms having been dropped in the derivations.



## 162 Factors Which Affect the Correlation Coefficient

Let  $I = X_1/X_2$ ; then it can be shown that the mean and standard deviation of such an index or ratio will be approximately

$$M_I = \frac{M_1}{M_2} (1 - r_{12}v_1v_2 + v_2^2) \quad (57)$$

$$\sigma_I = \frac{M_1}{M_2} \sqrt{v_1^2 - 2r_{12}v_1v_2 + v_2^2} \quad (58)$$

If we have 4 variables, the following formula for the correlation of indexes will yield a good approximation:

$$r \frac{X_1 X_2}{X_3 X_4} = \frac{r_{12}v_1v_2 - r_{14}v_1v_4 - r_{23}v_2v_3 + r_{34}v_3v_4}{\sqrt{v_1^2 + v_3^2 - 2r_{13}v_1v_3} \sqrt{v_2^2 + v_4^2 - 2r_{24}v_2v_4}} \quad (59)$$

Although these formulas are very useful for determining means, sigmas, and the correlations for ratios in terms of means, sigmas, and correlation coefficients for the original variables, their use is somewhat limited in that generally one cannot know whether the index distribution is normal, nor can one make a statement concerning linearity and homoscedasticity for the correlation between 2 indexes. Such information, if needed, must be obtained by first determining the numerical value of the indexes for each individual and then making distributions.

Several special cases can be deduced from formula (59). Thus the correlation between  $X_1/X_3$  and  $X_2$  is exactly equivalent to that between  $X_1/X_3$  and  $X_2/1$ ; i.e.,  $X_4$  is set equal to 1, which makes  $v_4 = 0$ , and therefore all terms in (59) involving the subscript 4 vanish. The correlation between  $X_1/X_3$  and the reciprocal of a variable would be obtained by setting  $X_2 = 1$ , i.e., letting  $1/X_4$  be the reciprocal; then  $v_2 = 0$ , whence the desired formula can be obtained by dropping all terms involving  $v_2$ . Likewise the correlation can be deduced for  $1/X_3$  with  $1/X_4$ , for  $1/X_3$  with  $X_2$ , and for  $X_1/X_3$  with  $X_2/X_3$ . This last correlation is of particular interest because it is possible to find a relationship between these 2 indexes even though the 3 original variables are uncorrelated.

By substituting  $X_3$  for  $X_4$ , i.e., replacing subscript 4 by 3, an expression for the correlation of indexes having a common variable denominator can readily be obtained. It will be

$$\frac{r_{\frac{X_1}{X_3} \frac{X_2}{X_3}}}{\frac{X_1}{X_3} \frac{X_2}{X_3}} = \frac{r_{12}v_1v_2 - r_{13}v_1v_3 - r_{23}v_2v_3 + v_3^2}{\sqrt{v_1^2 + v_3^2} \sqrt{v_2^2 + v_3^2} - 2r_{13}v_1v_3 \sqrt{v_2^2 + v_3^2} - 2r_{23}v_2v_3} \quad (60)$$

If  $r_{12} = r_{13} = r_{23} = 0$ , this becomes

$$\frac{v_3^2}{\sqrt{v_1^2 + v_3^2} \sqrt{v_2^2 + v_3^2}}$$

and if the  $v$ 's are equal, the value of the index correlation will be .50 even though there is no relationship between the original variables. This is known as *spurious correlation* due to indexes. There are instances, however, in which an analysis of the interrelations of ratios is of just as much import as the analysis of the variables from which the indexes are obtained, and therefore it does not follow that the correlation between ratios having a common denominator is necessarily misleading.

It has been asserted that the correlation between IQ's derived from 2 tests or 2 forms of the same test will be spuriously high because of the common variable denominator, age. It can be shown, however, that such a correlation will not be spurious unless the 2 sets of IQ's are correlated with age. If the IQ-vs.-age correlations are both positive or both negative, the index correlation will be spuriously high; if one is negative and the other positive, spuriously low. Thus, rather than make a blanket statement to the effect that the correlation between IQ's is spuriously high, we should say that it can be spuriously high or low or not spurious at all, according to the IQ-vs.-age correlations. It should be remembered that, even though the IQ's based on an ideal (properly constructed and standardized) test will be uncorrelated with age, a nonzero relationship might be produced for a single school-grade group by the selective factors that operate in age-grade location. Within a single grade group in a school system where acceleration is permitted, the younger children are likely to be the brighter, i.e., have the higher IQ's, thus producing negative correlations for sets of IQ's with age, and consequently a spuriously high correlation between IQ's.

## PART-WHOLE CORRELATION

Another type of spurious correlation arises when a total score is correlated with a subscore which is a part of the total score. Suppose that a total score is made up of 3 parts,  $X_t = X_1 + X_2 + X_3$ , and that we correlate  $X_1$  against  $X_t$ . Ordinarily in such situations the components will themselves be correlated positively. It should be obvious that the extent to which  $X_1$  correlates with  $X_t$  is more or less dependent upon the fact that  $X_t$  includes  $X_1$ . It does not follow, however, that a high value for  $r_{1t}$  is not meaningful, even though spurious. For instance, a high value for  $r_{1t}$  would, regardless of spuriousness, justify the use of  $X_1$  in lieu of the battery of 3 subtests. There are times when one may wish to know how highly a subtest correlates with a total, based on any number of parts, *minus* the subtest. This correlation is given by

$$r_{1(t-1)} = \frac{r_{1t}\sigma_t - \sigma_1}{\sqrt{\sigma_t^2 + \sigma_1^2 - 2r_{1t}\sigma_1\sigma_t}} \quad (61)$$

## HETEROGENEITY WITH RESPECT TO A THIRD VARIABLE

We have already discussed the influence on  $r$  of heterogeneity with regard to one or both the variables being correlated. Suppose variables  $X_1$  and  $X_2$  are 2 different traits, each of which is related to age as the third variable. Then an older individual will tend to be higher on both tests than a younger individual. In other words heterogeneity with respect to age will tend to produce correlation between  $X_1$  and  $X_2$ , and our present problem is to develop a method for correcting  $r_{12}$  so that we can estimate what the correlation between  $X_1$  and  $X_2$  would be if age were constant.

Suppose  $r_{12}$ ,  $r_{13}$ ,  $r_{23}$ , and the several means and standard deviations are known; then let us visualize the 3 scatter diagrams. The scatter for  $r_{12}$  will be somewhat elongated as a result of the influence of age, since variation in both  $X_1$  and  $X_2$  are here supposed to be partly due to age variation. What is needed is the correlation, between measures of  $X_1$  and  $X_2$ , which has been freed from the influence of age. If we were to express each  $X_1$  in the first array of the scatter for  $r_{13}$  as a deviation from the mean of this array and were to do the same for all other  $X_1$ 's in the scatter—each as a deviation from the mean of the array in which it falls—we would have scores expressed as deviations from the means of

the several ages. These deviations will be independent of age. As an example, suppose an 8-year-old individual scores 28 and the mean of 8-year-olds is 25, and a 14-year-old individual scores 54 and the mean of 14-year-olds is 51. The second individual scores higher than the first because he is older, but each would have a deviation (from his own age mean) of plus 3. Obviously, if we also expressed the  $X_2$  scores as deviations from the averages for the several ages, they too would be independent of age influences. Now, if we correlated these deviations (from age means) we would be correlating sets of  $X_1$  and  $X_2$  scores which would be free from age, and hence we would arrive at a correlation, between variables  $X_1$  and  $X_2$ , which would not be affected by age heterogeneity.

**Partial correlation.** The task of determining the correlation between 2 variables, with the influence of a third eliminated, can always be accomplished by actually computing all the deviations and then making a scatter diagram from which the  $r$  can be determined; but, in those cases in which we can assume linearity of regression for  $X_1$  on  $X_3$  and  $X_2$  on  $X_3$ , it is possible to set up a method for determining the desired correlation from the 3 correlation coefficients between the 3 variables. If linearity exists, we can correlate the deviations from the 2 regression lines instead of from the array means (or means for several ages if age is the third variable). Since

$$x'_1 = r_{13} \frac{\sigma_1}{\sigma_3} x_3 \quad \text{and} \quad x'_2 = r_{23} \frac{\sigma_2}{\sigma_3} x_3$$

the 2 sets of deviation-from-regression scores will be

$$x_1 - x'_1 = x_1 - r_{13} \frac{\sigma_1}{\sigma_3} x_3 \quad \text{and} \quad x_2 - x'_2 = x_2 - r_{23} \frac{\sigma_2}{\sigma_3} x_3$$

The correlation of these deviation scores, which is designated by the symbol  $r_{12 \cdot 3}$  (read: the correlation between  $X_1$  and  $X_2$  with  $X_3$  held constant) and known as the *partial correlation coefficient*, becomes

$$\begin{aligned} r_{12 \cdot 3} &= \frac{\Sigma(x_1 - x'_1)(x_2 - x'_2)}{N\sigma_{x_1 - x'_1}\sigma_{x_2 - x'_2}} \\ &= \frac{\Sigma\left(x_1 - r_{13} \frac{\sigma_1}{\sigma_3} x_3\right)\left(x_2 - r_{23} \frac{\sigma_2}{\sigma_3} x_3\right)}{N\sigma_{x_1 - x'_1}\sigma_{x_2 - x'_2}} \end{aligned}$$

## 166 Factors Which Affect the Correlation Coefficient

Multiplying and summing the numerator, and noting that the  $\sigma$ 's in the denominator are nothing more than the errors of estimate,  $\sigma_{1.3}$  and  $\sigma_{2.3}$ , we have

$$r_{12.3} = \frac{\sum x_1 x_2 - r_{23} \frac{\sigma_2}{\sigma_3} \sum x_1 x_3 - r_{13} \frac{\sigma_1}{\sigma_3} \sum x_2 x_3 + r_{13} r_{23} \frac{\sigma_1 \sigma_2}{\sigma_3^2} \sum x^2_3}{N \sigma_1 \sqrt{1 - r_{13}^2} \sigma_2 \sqrt{1 - r_{23}^2}}$$

Dividing by  $N$ , cancelling  $\sigma$ 's, and collecting like terms, we get

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{1 - r_{13}^2} \sqrt{1 - r_{23}^2}} \quad (62)$$

This formula definitely assumes the linearity of the 2 regression lines for predicting  $X_1$  and  $X_2$  from  $X_3$ . Whether we correlate deviations from array means or use formula (62), we end with a correlation which has been freed of the influence of the third, or eliminated, variable. If, for example, age is the third variable, the partial correlation coefficient represents an estimate of what the correlation would be if we held age constant by the use of individuals of *any* one of the several age levels present in the original group.

The difference between  $r_{12.3}$  and  $r_{12}$  indicates how much of the correlation between variables 1 and 2 is due to the influence of heterogeneity of a third variable. Obviously, if the third variable is unrelated to  $X_1$  and  $X_2$ , the partial  $r$  will equal  $r_{12}$ , and if either  $r_{13}$  or  $r_{23}$  is negative and  $r_{12}$  positive, "partialing out"  $X_3$  will raise the correlation. Is this reasonable?

The difficulties encountered in determining the direction of causation make it necessary to be careful in the use of the partial correlation technique. When it is said that heterogeneity with respect to a third variable ( $X_3$ ) has in part (or entirely) produced correlation between  $X_1$  and  $X_2$ , one must ask how the influence of  $X_3$  comes about. Now if it can be argued that variation in  $X_3$  is a cause of variation in  $X_1$  and  $X_2$ , it is readily seen that  $r_{12}$  is at least in part attributable to the fact that  $X_1$  and  $X_2$  have a common source of variation. The partial,  $r_{12.3}$ , tells us the degree of correlation between  $X_1$  and  $X_2$  which would exist provided variation in  $X_3$  were controlled. But if it cannot be claimed that  $X_3$  produces variation in  $X_1$  and  $X_2$ , the interpretation of the partial  $r$  is far from clear. Suppose  $X_1$  precedes  $X_3$  in a temporal

sense so that we know variation on  $X_3$  couldn't possibly contribute to variation in  $X_1$ , does it make sense to interpret  $r_{12.3}$  as the correlation between  $X_1$  and  $X_2$  with the influence of  $X_3$  nullified when we know that  $X_3$  could not influence  $X_1$ ? Stated differently, the only way that  $X_3$  can produce or contribute to the correlation between  $X_1$  and  $X_2$  is by way of  $X_3$  producing variation in  $X_1$  and  $X_2$ .

The technique can be extended for "partialing out" or eliminating more than 1 variable. Thus, to obtain an estimate of  $r_{12}$  with  $X_3$  and  $X_4$  held constant, we can use

$$r_{12.34} = \frac{r_{12.4} - r_{13.4}r_{23.4}}{\sqrt{1 - r_{13.4}^2} \sqrt{1 - r_{23.4}^2}}$$

which is in terms of first-order partials calculable by formula (62).

The sampling error of the partial coefficient may be handled by the  $z$  transformation. The standard error of the corresponding  $z$  will be  $1/\sqrt{N-4}$  when only 1 variable has been eliminated, and  $1/\sqrt{N-5}$  when 2 variables have been eliminated.

The partial correlation coefficient based on a small sample can also be tested for significance by the  $t$  technique. If 1 variable has been eliminated, we have

$$t = \frac{r_{12.3}}{\sqrt{\frac{1 - r_{12.3}^2}{N-3}}}$$

with  $df = N - 3$ . An additional degree of freedom is lost for each additional variable eliminated.

A perplexing and often-recurring question with regard to the interrelations of 3 variables is this: Are the correlations consistent among themselves, or, if  $r_{12}$  and  $r_{13}$  are known, what are the possible limits for  $r_{23}$ ? If  $r_{12} = \text{unity}$  and  $r_{13} = \text{unity}$ ,  $r_{23}$  must also equal unity, but, if  $r_{12} = 0$  and  $r_{13} = 0$ , does it follow that  $r_{23} = 0$ ? It can be shown that the limits for the correlation  $r_{23}$  will always be  $r_{12}r_{13} \pm \sqrt{1 - r_{12}^2 - r_{13}^2 + r_{12}^2 r_{13}^2}$ .

#### EXAMPLES:

When  $r_{12}$  and  $r_{13}$  each equal .90, the limits for  $r_{23}$  are +.62 and +1.00;  
 " " " " " " .50, " " " " " " -.50 and +1.00;  
 " " " " " " .25, " " " " " " -.875 and +1.00.



**SUMMARY**

In this chapter, consideration has been given to factors which have a bearing on the magnitude of the correlation coefficient. If any of these is operative in the case of a particular coefficient, it is the responsibility of the investigator to qualify his conclusions accordingly. Published reports of correlational studies should include:

1. A definition of the population being sampled and a statement of the method used in drawing the sample.

2. The size of the sample and an adequate treatment of sampling by means of nonantiquated formulas.

3. The means and particularly the standard deviations of the variables being correlated, with some indication as to whether the sample is typical as regards heterogeneity with respect to the variables under consideration.

4. The reliability coefficients for the measures and the method of determining reliability.

5. A statement relative to the homogeneity of the sample with respect to possibly relevant variables such as age, sex, race.

6. A defense or precise interpretation of any reported correlations involving indexes or of any part-whole correlations.

The researcher who is cognizant of the assumptions requisite for a given interpretation of a correlation coefficient and who is also fully aware of the many factors which may affect its magnitude will not regard the correlational technique as an easy road to scientific discovery.

## Multiple Correlation

So far our discussion of correlation has been concerned chiefly with the prediction of one variable from another or the attributing of a portion of the variance of one variable to the action of a second variable. We shall next consider the case where it is desired to predict one variable by using several other variables as a team of predictors, or where, if causation can be assumed, an attempt is made to analyze the variance for one variable into components or parts attributable to the action of two or more other variables. There is a close connection between the predicting and the analyzing problems; let us first consider the method of predicting one variable on the basis of other variables.

### THE THREE-VARIABLE PROBLEM

For simplicity, consider the problem of predicting  $X_1$  from a knowledge of  $X_2$  and  $X_3$ . The  $X_1$  variable is frequently called the criterion, or dependent variable. If we had  $X_1$  to be predicted from  $X_2$  alone, we would have exactly the same situation as predicting  $Y$  from  $X$ . That is, the linear prediction equation (in gross score form)

$$Y' = BX + A$$

becomes

$$X'_1 = BX_2 + A$$

and the deviation form

$$y' = bx + a$$

becomes

$$x'_1 = bx_2 + a$$

It will be recalled that the values of the constants,  $B$  and  $A$ , or  $b$  and  $a$ , were so determined as to give the maximum predictability,

and that  $B$  and  $A$  turned out to be functions of the correlation coefficient between the two variables and of the means and standard deviations for the variables. The equation which resulted from giving  $A$  and  $B$  specific values was said to be the equation of the best-fitting line—the error of prediction was minimized.

Now, if we wish to predict  $X_1$  from  $X_2$  and  $X_3$ , we start with an equation of the form

$$X'_1 = B_2X_2 + B_3X_3 + A \quad (63)$$

which can be written in deviation units as

$$x'_1 = b_2x_2 + b_3x_3 + a$$

Either of these forms represents the equation of a plane. It can be shown that  $B_2 = b_2$  and  $B_3 = b_3$ . In fact, this is rather obvious when we consider the meaning of these  $B$  or  $b$  coefficients. They represent the slope of the plane;  $B_2$  is the slope which the plane makes with the  $x_2$  axis, and  $B_3$  the slope with regard to the  $x_3$  axis. When we shift from raw to deviation scores, we are merely shifting the origin, or point of reference, to the intersection of the means, and this point in terms of deviation scores becomes zero. This shift of the frame of reference does not change the position or angle of the plane; hence  $B_2 = b_2$  and  $B_3 = b_3$ . (The student will recall that, for the ordinary two-variable problem, the slope of the line was equal to  $B$  or  $b$ .)

It remains to attach meaning to  $A$  and  $a$ . In the equation  $Y' = BX + A$ , it was noted that the constant  $A$  was the  $Y$  intercept, i.e., the value of  $Y$  where the line cut the  $y$  axis. It was also found that  $a = 0$ ; i.e., that in the deviation form the line cut the  $y$  axis at the origin. Perhaps the student has already anticipated, by analogy, that the  $A$  in our three-variable equation is the value of  $X_1$  where the plane cuts the  $x_1$  axis, and that the value of  $a$  will become zero.

Before going farther, it might be well to take a look at the problem geometrically. In the case of two variables, after plotting the  $X$  and  $Y$  values in a scattergram, we can readily picture the meaning of  $B$  and  $A$ , and also obtain some notion of why certain values of  $B$  and  $A$  will lead to better predictions than those obtained by other values. In the case of three variables,  $X_1$ ,  $X_2$ , and  $X_3$ , we have a trio instead of a pair of measurements. In

order to draw up a plot of  $N$  such sets of measurements, we will need to use a three-dimensional scheme. Instead of placing a tally mark in a cell defined by an interval along the  $x$  axis and one along the  $y$  axis, we now have to consider a cell as defined by intervals on the  $x_1$ , the  $x_2$ , and the  $x_3$  axes. Instead of a square cell, we have a cubical cell.

Suppose an individual's three scores fall in intervals  $i_1$ ,  $i_2$ , and  $i_3$ ; then his "tally" will be placed in the cubicle formed at the intersection of these three intervals. The total number of cubicles will be the product of the number of intervals on each axis, and an individual's location in the "box" will depend upon all three of his scores. The student may be at a loss to know just how one could make such a three-dimensional scattergram. Actually, this diagram is not necessary, but it is of interest to imagine what such a three-way distribution would look like. If the correlations,  $r_{12}$ ,  $r_{13}$ , and  $r_{23}$ , are fairly high (and positive), and if we think of the frequencies in the several cubicles as being represented by dots (or different degrees of density), then the swarm of dots will extend from the lower left front to the upper right back of the box. The greatest density will be at the center of this swarm, and the density or frequency will fall off in all directions from the center. The swarm will have the general shape and appearance of a watermelon (ellipsoidal).

Imagine that a plane is to be cut through this swarm. Our job is to so locate the plane that, when we start upward vertically from any point on the bottom of the box, say the spot defined by any pair of values for  $X_2$  and  $X_3$ , we will find that the altitude, i.e., the distance along the  $x_1$  axis at which the plane is reached, will constitute the best estimate of  $X_1$  for individuals having any given  $X_2$  and  $X_3$  scores. With a little reflection, the reader can see that, of many ways of placing the plane, some positions will obviously give very poor estimates, whereas others will lead to better estimates. What we need is that plane which, for the given  $N$  sets of  $X_1$ ,  $X_2$ , and  $X_3$  scores, will yield the best possible estimates.

The criterion of "best" is a least square affair—the sum of the squares of the errors of estimate shall be a minimum. The task is really that of determining the values of  $A$ ,  $B_2$ , and  $B_3$  in formula (63) so that

$$\Sigma(X_1 - X'_1)^2$$

is a minimum. That is, we are to assign to  $A$ ,  $B_2$ , and  $B_3$  those values which will permit the best possible estimate of an unknown  $X_1$  when we know the  $X_2$  and  $X_3$  values for the individual. The principle to be used is exactly the same as that employed to obtain the optimum value for  $B$  and  $A$  for the two-variable problem, but the present problem is more complicated because we have to determine the values for three constants.

**Derivation of regression equations.** Our task is simplified if deviation scores are used, and we assume  $a = 0$  (if we carried  $a$  along, it would prove to be zero). It is simplified somewhat more if we transform all three sets of scores into standard score form, i.e., if we set  $z = (X - M)/\sigma$ . Then our equation becomes

$$z'_1 = \beta_2 z_2 + \beta_3 z_3 \quad (64)$$

It should be noted that, since we are changing the size of our unit of measure, it cannot be argued that  $\beta_2$  will equal  $B_2$  or  $b_2$ . The task now is to determine the value of the *beta coefficients*,  $\beta_2$  and  $\beta_3$ , so as to have the best possible estimate of  $z_1$ , or so that the average of the squared errors, or

$$\frac{1}{N} \sum (z_1 - z'_1)^2$$

shall be a minimum. Since  $z_1 - z'_1 = z_1 - \beta_2 z_2 - \beta_3 z_3$ , the function,  $f$ , to be minimized is

$$f = \frac{1}{N} \sum (z_1 - \beta_2 z_2 - \beta_3 z_3)^2$$

To determine the values of  $\beta_2$  and  $\beta_3$  which will make this function a minimum, use is made of the calculus. We take the partial derivative of the function first with respect to  $\beta_2$ , then with respect to  $\beta_3$ . Thus,

$$\frac{\partial f}{\partial \beta_2} = \frac{-2 \sum z_2}{N} (z_1 - \beta_2 z_2 - \beta_3 z_3)$$

$$\frac{\partial f}{\partial \beta_3} = \frac{-2 \sum z_3}{N} (z_1 - \beta_2 z_2 - \beta_3 z_3)$$

These two derivatives are to be set equal to zero and then solved simultaneously for the two unknowns,  $\beta_2$  and  $\beta_3$ . Performing the

indicated multiplications, summing, and dividing each equation by 2, we get

$$-\frac{\Sigma z_1 z_2}{N} + \beta_2 \frac{\Sigma z_2^2}{N} + \beta_3 \frac{\Sigma z_2 z_3}{N} = 0$$

$$-\frac{\Sigma z_1 z_3}{N} + \beta_2 \frac{\Sigma z_2 z_3}{N} + \beta_3 \frac{\Sigma z_3^2}{N} = 0$$

Since we are dealing with standard scores, we can now capitalize on certain properties thereof, namely, that the sum of their squares divided by  $N$  is unity, while any sum of cross products divided by  $N$  is the correlation between the two variables involved in the cross products. Thus, we have

$$-r_{12} + \beta_2 + \beta_3 r_{23} = 0$$

$$-r_{13} + \beta_2 r_{23} + \beta_3 = 0$$

or

$$\beta_2 + r_{23}\beta_3 - r_{12} = 0 \tag{65}$$

$$r_{23}\beta_2 + \beta_3 - r_{13} = 0$$

Since the  $r$ 's in the equations are determinable for any given sample of data, they are in effect knowns, whereas the  $\beta$ 's are unknowns. We therefore have two simultaneous equations with two unknowns. These can readily be solved by a number of methods which the student will find in an algebra textbook. Straightforward solution gives

$$\beta_2 = \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}$$

$$\beta_3 = \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}$$

As soon as we have computed the  $r$ 's, we can easily determine the  $\beta$ 's. The obtained numerical values can then be substituted in the prediction equation

$$z'_1 = \beta_2 z_2 + \beta_3 z_3$$

so that for a given pair of  $z_2$  and  $z_3$  values we can predict the standard score on the criterion variable. However, in practice it is ordinarily more convenient to deal with raw scores; hence we need



our prediction equation in raw score form. Obviously, if we replace the  $z$ 's in the above equation by their values in terms of raw scores, means, and standard deviations, we will have

$$\frac{X'_1 - M_1}{\sigma_1} = \beta_2 \frac{X_2 - M_2}{\sigma_2} + \beta_3 \frac{X_3 - M_3}{\sigma_3}$$

or

$$\frac{X'_1}{\sigma_1} - \frac{M_1}{\sigma_1} = \beta_2 \frac{X_2}{\sigma_2} - \beta_2 \frac{M_2}{\sigma_2} + \beta_3 \frac{X_3}{\sigma_3} - \beta_3 \frac{M_3}{\sigma_3}$$

Multiplying by  $\sigma_1$  and rearranging terms, we have

$$X'_1 = \beta_2 \frac{\sigma_1}{\sigma_2} X_2 + \beta_3 \frac{\sigma_1}{\sigma_3} X_3 + \left( M_1 - \beta_2 \frac{\sigma_1}{\sigma_2} M_2 - \beta_3 \frac{\sigma_1}{\sigma_3} M_3 \right) \quad (66)$$

from which we see that our original  $B_2$  must equal  $\beta_2(\sigma_1/\sigma_2)$ ,  $B_3 = \beta_3(\sigma_1/\sigma_3)$ , and  $A =$  the parentheses term. Thus we can readily determine the numerical values of  $B_2$ ,  $B_3$ , and  $A$  and thereby have the constants for the prediction equation. Actually, the values of  $B_2$  and  $B_3$  are the optimum weights to be assigned to  $X_2$  and  $X_3$  in order to predict  $X_1$ .

**Error of estimate.** The accuracy of the prediction of  $X_1$  by the best combination of  $X_2$  and  $X_3$  can be ascertained by examining the error term, i.e.,  $X_1 - X'_1$  or  $\sigma_1(z_1 - z'_1)$ . The sum of the squares for the errors divided by  $N$  will yield the variance of the errors. The square root would correspond to the standard error of estimate. Let  $\sigma_{z_1 \cdot z_2}$  be this error (in sigma units), then

$$\begin{aligned} \sigma_{z_1 \cdot z_2}^2 &= \frac{\sum (z_1 - z'_1)^2}{N} \\ &= \frac{\sum (z_1 - \beta_2 z_2 - \beta_3 z_3)^2}{N} \\ &= \frac{\sum z_1^2}{N} + \beta_2^2 \frac{\sum z_2^2}{N} + \beta_3^2 \frac{\sum z_3^2}{N} - \frac{2\beta_2 \sum z_1 z_2}{N} - \frac{2\beta_3 \sum z_1 z_3}{N} \\ &\quad + \frac{2\beta_2 \beta_3 \sum z_2 z_3}{N} \\ &= 1 + \beta_2^2 + \beta_3^2 - 2\beta_2 r_{12} - 2\beta_3 r_{13} + 2\beta_2 \beta_3 r_{23} \end{aligned}$$

which by algebraic manipulation reduces to

$$\sigma^2_{z_1 \cdot z_3} = 1 - (\beta_2 r_{12} + \beta_3 r_{13}) \quad (67)$$

in terms of standard scores. Then  $\sigma^2_{z_1}$  times this would give the error variance for raw scores.

**Multiple  $r$ .** We next define the *multiple correlation coefficient* as the correlation between  $z_1$  and the best estimate of  $z_1$  from a knowledge of  $z_2$  and  $z_3$ . In symbols,

$$\begin{aligned} r_{1 \cdot 23} &= r_{z_1 z'_{11}} = \frac{\Sigma z_1 z'_{11}}{N \sigma_{z_1} \sigma_{z'_{11}}} \\ &= \frac{\Sigma z_1 (\beta_2 z_2 + \beta_3 z_3)}{N \sigma_{z'_{11}}} \end{aligned} \quad (68)$$

Note that, although  $\sigma_{z_1} = 1$ , it does not follow that  $\sigma_{z'_{11}} = 1$ . In order to evaluate this last  $\sigma$ , we write

$$z_1 = z'_{11} + z_{1 \cdot 23}$$

That is, we think of  $z_1$  as being made up of two parts, that which we can estimate plus a residual. It can easily be shown that these two parts are independent of each other; hence by the variance theorem we have

$$\sigma^2_{z_1} = \sigma^2_{z'_{11}} + \sigma^2_{z_{1 \cdot 23}}$$

or

$$1 = \sigma^2_{z'_{11}} + \sigma^2_{z_{1 \cdot 23}}$$

then

$$\sigma^2_{z'_{11}} = 1 - \sigma^2_{z_{1 \cdot 23}}$$

But  $\sigma^2_{z_{1 \cdot 23}}$  is nothing more than the variance of the prediction errors as given by (67); therefore

$$\sigma_{z'_{11}} = \sqrt{\beta_2 r_{12} + \beta_3 r_{13}}$$

Then, by substituting in formula (68), we have

$$\begin{aligned} r_{1 \cdot 23} &= \frac{\Sigma z_1 (\beta_2 z_2 + \beta_3 z_3)}{N \sqrt{\beta_2 r_{12} + \beta_3 r_{13}}} \\ &= \frac{\beta_2 \Sigma z_1 z_2 + \beta_3 \Sigma z_1 z_3}{N \sqrt{\beta_2 r_{12} + \beta_3 r_{13}}} = \frac{\beta_2 r_{12} + \beta_3 r_{13}}{\sqrt{\beta_2 r_{12} + \beta_3 r_{13}}} \\ &= \sqrt{\beta_2 r_{12} + \beta_3 r_{13}} \end{aligned} \quad (69)$$

We thus see that, as soon as the  $\beta$ 's are determined, we can write the regression equation for predicting  $z_1$  from  $z_2$  and  $z_3$  and can also specify the degree of correlation and calculate the error of estimate. This error obviously can be written from formulas (67) and (69) as

$$\sigma_{1.23} = \sigma_1 \sqrt{1 - r_{1.23}^2} \quad (70)$$

which is in terms of raw scores.

Formula (70) has been used frequently to define the multiple correlation coefficient. Stated explicitly,

$$r_{1.23}^2 = 1 - \frac{\sigma_{1.23}^2}{\sigma_1^2} = 1 - \sigma_{\epsilon_{1.23}}^2$$

Then, by substituting from (67), we again arrive at (69).

The student will note the similarity of formula (70) to the ordinary error of estimate for the bivariate situation. Thus the multiple correlation coefficient can be interpreted, in terms of reduction in the error of estimate, in exactly the same manner as the ordinary bivariate correlation coefficient. The only difference is that we are now determining the regression coefficients, or weights for two variables as a team, so as to get the best possible prediction of a third variable, whereas in the bivariate situation only one regression coefficient is necessary. A multiple correlation coefficient of .60 has, aside from minor qualifications to be discussed later, the same meaning in a predictive sense as an ordinary correlation of .60. Furthermore, the interpretation in terms of contribution to variance also holds for the multiple correlation coefficient; i.e., if one can assume causation, it may be said that a multiple  $r$  of .60 indicates that 36 per cent of the variance in the criterion or dependent variable can be attributed to variation in the two independent variables.

**Relative weights.** The question arises as to the relative importance of the two variables as contributors to variation in the criterion variable. The  $B$  coefficients in the regression equation have, at times, been misinterpreted as indicating the relative contribution of the two independent variables. The reader need only be reminded that the two  $B$  coefficients usually involve different units of measurement (one may be in terms of feet and the other in pounds); hence they are not comparable at all. If  $B_2$  is numerically twice  $B_3$ , it does not follow that  $X_2$  is twice as im-

portant as  $X_3$ . In order to get around this difficulty, we must think in terms of standard scores; these will be comparable, and hence the  $\beta$  coefficients in the standard score form of the regression equation will be comparable.

Since

$$\sigma^2_{z_1} = \sigma^2_{z'_1} + \sigma^2_{z_1 \cdot z_2}$$

or

$$1 = \sigma^2_{z'_1} + \sigma^2_{z_1 \cdot z_2}$$

and

$$1 - \sigma^2_{z_1 \cdot z_2} = r^2_{1 \cdot 23}$$

it follows that

$$r^2_{1 \cdot 23} = \sigma^2_{z'_1}$$

That is,  $r^2_{1 \cdot 23}$ , which corresponds to the percentage of variance explained, is equal to  $\sigma^2_{z'_1}$ , or the variance of the predicted standard scores. This variance could be determined by actually making  $N$  predictions of  $z_1$  from the  $N$  pairs of values of  $z_2$  and  $z_3$  and then computing the sigma for the distribution of these predicted values. This is not done in practice, since the value of this sigma squared is  $r^2_{1 \cdot 23}$ , which is easily calculated once the  $\beta$ 's have been determined.

But note that, since

$$z'_1 = \beta_2 z_2 + \beta_3 z_3$$

we can indicate the value of  $\sigma^2_{z'_1}$  as

$$\begin{aligned} \sigma^2_{z'_1} &= \frac{\Sigma(z'_1)^2}{N} = \frac{\Sigma(\beta_2 z_2 + \beta_3 z_3)^2}{N} \\ &= \frac{\beta_2^2 \Sigma z_2^2 + \beta_3^2 \Sigma z_3^2 + 2\beta_2 \beta_3 \Sigma z_2 z_3}{N} \end{aligned}$$

which becomes

$$\sigma^2_{z'_1} = \beta_2^2 + \beta_3^2 + 2\beta_2 \beta_3 r_{23} \quad (71)$$

In other words, the predicted variance, which corresponds to the "explained" variance, can be broken down into three additive components. We thus see that the relative importance of the variables  $X_2$  and  $X_3$  in "explaining" or "causing" variation in  $X_1$  can be judged by the magnitude of the squares of the  $\beta$  coefficients. The third term in formula (71) represents a joint contribution which, it will be seen, is a function of the amount of correlation between the two predicting variables.

Summarizing, it can be said that the fundamental problem in multiple correlation is that of obtaining the optimum weighting to be assigned to independent variables ( $X_2$  and  $X_3$ ) in predicting or explaining variation in a dependent variable,  $X_1$ . That is, we determine the value of  $B_2$ ,  $B_3$ , and  $A$  in the equation

$$X'_1 = B_2X_2 + B_3X_3 + A$$

so as to get the best possible estimate of  $X_1$ . This is resolved by working with the prediction equation in standard score form with  $\beta$  coefficients. The value of each  $\beta$  is determinable from the inter-correlations among the three variables. Once the  $\beta$ 's are calculated, we can: (1) readily compute the  $B$  coefficients needed in the raw score form of the prediction equation; (2) determine the value of the multiple correlation coefficient and the error of estimate; (3) ascertain the relative importance of the independent variables as predictors or, if causation can be assumed, as contributors to the variance of the dependent or criterion variable. It is important to note that the multiple correlation coefficient represents the maximum correlation to be expected between the dependent variable and a linearly additive combination of  $X_2$  and  $X_3$ .

### MORE THAN THREE VARIABLES

Suppose that we have a dependent variable and four independent variables which might be used as predictors or which might be thought of as causes of variation in the dependent variable. The cause and effect, as opposed to concomitant, relationship among variables is a logical problem which must be faced by the investigator as a logician rather than as a statistician. Whether one resorts to the multiple correlation technique as an aid in predicting or as an aid in analysis will depend entirely upon the problem being attacked; the mechanical solution is the same, but the investigator must choose the interpretation which best suits his purpose.

For a five-variable problem, we need the constants in the regression or prediction equation,

$$X'_1 = B_2X_2 + B_3X_3 + B_4X_4 + B_5X_5 + A$$

which can be written in standard score form as

$$z'_1 = \beta_2z_2 + \beta_3z_3 + \beta_4z_4 + \beta_5z_5$$

As in the three-variable situation, the problem is that of determining the optimum values of the  $B$ 's or the  $\beta$ 's so as to get the best possible prediction of  $X_1$  or  $z_1$ , i.e., so that

$$\frac{\Sigma(X_1 - X'_1)^2}{N}$$

or

$$\frac{\Sigma(z_1 - z'_1)^2}{N}$$

shall be as small as possible. The mathematical solution is easier by way of the standard score form of the regression equation. We have the function

$$f = \frac{\Sigma(z_1 - z'_1)^2}{N} = \frac{\Sigma(z_1 - \beta_2 z_2 - \beta_3 z_3 - \beta_4 z_4 - \beta_5 z_5)^2}{N} \quad (72)$$

which is to be minimized by assigning proper values to the  $\beta$ 's. These values are obtained by taking the derivative of the function with respect to, and in order for, each of the  $\beta$ 's. This will yield four derivatives which when set equal to zero will give us four equations involving the four unknown  $\beta$ 's. These equations can then be solved as simultaneous equations in order to determine the values of the  $\beta$ 's. The obtained  $\beta$ 's will be such that the sum of the squares of  $z_1 - z'_1$  will be the least possible; i.e., we will have the best possible estimate of  $z_1$  from an additive combination of the four independent variables.

The student of the calculus can readily verify that the four equations obtained by taking derivatives of formula (72) will take the following form (when set equal to zero):

$$\left. \begin{aligned} \beta_2 + \beta_3 r_{23} + \beta_4 r_{24} + \beta_5 r_{25} - r_{12} &= 0 \\ \beta_2 r_{23} + \beta_3 + \beta_4 r_{34} + \beta_5 r_{35} - r_{13} &= 0 \\ \beta_2 r_{24} + \beta_3 r_{34} + \beta_4 + \beta_5 r_{45} - r_{14} &= 0 \\ \beta_2 r_{25} + \beta_3 r_{35} + \beta_4 r_{45} + \beta_5 - r_{15} &= 0 \end{aligned} \right\} \quad (73)$$

These equations result from steps exactly parallel to those used for the three-variable problem. The four  $\beta$ 's are unknowns, whereas, for any given batch of data, the  $r$ 's take on specific numerical values.



The extension of multiple correlation to include any number of variables involves the same principles as utilized here for the three- and the five-variable problem. For  $n$  variables, formula (64) becomes

$$z'_1 = \beta_2 z_2 + \beta_3 z_3 + \cdots + \beta_n z_n \quad (64a)$$

The extension of (66) as the gross score equation should be obvious. Formula (69) for the multiple correlation coefficient becomes

$$r_{1.23 \dots n} = \sqrt{\beta_2 r_{12} + \beta_3 r_{13} + \cdots + \beta_n r_{1n}} \quad (69a)$$

To solve for the unknown  $\beta$ 's, the student may resort to any of the schemes given in algebra textbooks for solving simultaneous equations. One method is by way of determinants and Cramer's rule. The coefficients of the unknowns are the intercorrelations among the four independent variables, whereas the constants in these equations are the respective correlations of the dependent with the independent variables. In the application of Cramer's rule, these constants are thought of as being on the right-hand side of the equation, i.e., shifted to the right of the equality mark, with the consequent change of sign. The student should keep in mind, however, the fact that the original sign of any of the computed correlation coefficients must be considered.

Solution by Cramer's rule becomes quite tedious and burdensome for a problem involving more than four or five variables. Indeed, this determinantal solution is practically impossible for problems involving a large number of variables. Fortunately, there is available a simplified solution, but before turning to it, we would like to indicate some algebraic manipulations in terms of determinants.

It will be noted from the above simultaneous equations that all the intercorrelations among the five variables are involved. One can conveniently arrange these correlations in a table, or in determinantal form. Thus we can define a major determinant as

$$D = \begin{vmatrix} 1 & r_{12} & r_{13} & r_{14} & r_{15} \\ r_{12} & 1 & r_{23} & r_{24} & r_{25} \\ r_{13} & r_{23} & 1 & r_{34} & r_{35} \\ r_{14} & r_{24} & r_{34} & 1 & r_{45} \\ r_{15} & r_{25} & r_{35} & r_{45} & 1 \end{vmatrix}$$

If we were to delete the first row and first column, the minor which remains would involve the intercorrelations among the four independent variables. This minor might be conveniently symbolized as  $D_{11}$ ; i.e., we have deleted the column and the row which involves the subscript 1. If we were to delete the row which involves the subscript 1 and the column involving the subscript 2 throughout, we would symbolize the resulting minor as  $D_{12}$ .

Now it can be shown that

$$\beta_2 = \frac{D_{12}}{D_{11}}$$

or any  $\beta$ , say  $\beta_p$ , will be

$$\beta_p = (-1)^p \frac{D_{1p}}{D_{11}}$$

where the quantity  $(-1)^p$  is an indicator of either a positive or a negative sign, but the ultimate sign of  $\beta_p$  is also dependent upon whether the numerical values of the determinants are positive or negative. It can also be shown that the multiple correlation coefficient can be written as a function of determinants, thus

$$r_{1 \cdot 2345}^2 = 1 - \frac{D}{D_{11}}$$

The student who is interested in following a treatment of multiple correlation in terms of determinants is referred to T. L. Kelley's *Statistical method*.\*

## NUMERICAL SOLUTION

The solution of the simultaneous equations for the unknown  $\beta$ 's can best be accomplished by resort to Doolittle's method. This method is applicable to the solution of any simultaneous equations involving a major determinant which, like  $D$ , is symmetrical about the diagonal. It is also applicable to problems involving less or more than five variables. The first step is to write down the intercorrelations (coefficients of the unknown  $\beta$ 's) in the form indicated in Table 14, in which the right-hand column contains the correlation of each variable with the criterion or dependent variable. Negative signs are attached to these coefficients because, in essence,

\* Kelley, T. L., *Statistical method*, New York: Macmillan, 1924.

we are dealing with equations (73). Obviously, if the original sign of an  $r$  were negative, it would be preceded by a plus sign in an arrangement like that in Table 14.

Table 14. SCHEMA FOR ARRANGING  $r$ 's FOR DOOLITTLE SOLUTION

$X_2$	$X_3$	$X_4$	$X_5$	$X_1$
1	$r_{23}$	$r_{24}$	$r_{25}$	$-r_{12}$
	1	$r_{34}$	$r_{35}$	$-r_{13}$
		1	$r_{45}$	$-r_{14}$
			1	$-r_{15}$

As a numerical example, we shall use data from the Minnesota study of mechanical ability.† The sample size is 100.

Let  $X_1$  = Criterion (mechanical performance-quality).

$X_2$  = Minnesota assembling test.

$X_3$  = Minnesota spatial relations test.

$X_4$  = Paper form board.

$X_5$  = Interest analysis blank.

Since the several means and standard deviations will be needed, these are recorded in Table 15.

Table 15. MEANS AND  $SD$ 's (MINNESOTA DATA)

	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$
$M$	14.94	127.56	1422.90	46.60	107.00
$\sigma$	2.09	25.32	296.39	19.45	18.00

In Table 16 will be found the Doolittle solution for the  $\beta$  coefficients. Once these are known, the regression equation, in raw score form, can be written, and the multiple  $r$  and the error of estimate can be determined. The table includes an indication of the calculation of these values. The student will have to study carefully the schema of the Doolittle solution in order to grasp the necessary steps. We shall not attempt a complete exposition of the steps since the procedure of each step is indicated in the left-hand side of the table. A few remarks, however, will be of aid to the student.

† Paterson, D. G., *et al.*, *Minnesota mechanical ability tests*, Minneapolis: University of Minnesota Press, 1930.

Table 16. COMPUTATION OF MULTIPLE  $r$

	$X_2$	$X_3$	$X_4$	$X_5$	$X_1$	ck
(a)	1.00	.56	.49	.42	-.55	1.92
(b)		1.00	.63	.46	-.53	2.12
(c)			1.00	.39	-.52	1.99
(d)				1.00	-.64	1.63
(1): line (a)	1.00	.56	.49	.42	-.55	1.92
(2)	-1.00	-.56	-.49	-.42	.55	-1.92
(3): line (b)		1.000	.63	.46	-.53	2.12
(4): (1)(-.56)		-.314	-.274	-.235	.308	-1.075
(5): (3) + (4)		.686	.356	.225	-.222	1.045 ck
(6): (5)(-1/.686)		-1.000	-.519	-.328	.324	-1.524 ck
(7): line (c)			1.000	.39	-.52	1.99
(8): (1)(-.49)			-.240	-.206	.270	-.941
(9): (5)(-.519)			-.185	-.117	.115	-.542
(10): (7) + (8) + (9)			.575	.067	-.135	.507 ck
(11): (10)(-1/.575)			-1.000	-.116	.235	-.882 ck
(12): line (d)				1.000	-.64	1.63
(13): (1)(-.42)				-.176	.231	-.806
(14): (5)(-.328)				-.074	.073	-.343
(15): (10)(-.116)				-.008	.016	-.059
(16): (12 + (13) + (14) + (15)				.742	-.320	.422 ck
(17): (16)(-1/.742)				-1.000	.431	-.569 ck
Back solution						
From (17)				.431	$= \beta_5$	
From (11)			(.431)(-.116) + .235	$= \beta_4$	$= .185$	
From (6)		(.185)(-.519) + (.431)(-.328) + .324	$= \beta_3$	$= .087$		
From (2)	(.087)(-.56) + (.185)(-.49) + (.431)(-.42) + .55	$= \beta_2$	$= .230$			
Final checks						
(.230)(1.00) + (.087)(.56) + (.185)(.49) + (.431)(.42) - .55	$= .000$					
(.230)(.56) + (.087)(1.00) + (.185)(.63) + (.431)(.46) - .53	$= .001$					
(.230)(.49) + (.087)(.63) + (.185)(1.00) + (.431)(.39) - .52	$= .001$					
(.230)(.42) + (.087)(.46) + (.185)(.39) + (.431)(1.00) - .64	$= .000$					
From formula (66)						
$B_2 = (.230) \frac{2.09}{25.32} = .0190$	$B_3 = (.087) \frac{2.09}{296.39} = .0006$					
$B_4 = (.185) \frac{2.09}{19.45} = .0199$	$B_5 = (.431) \frac{2.09}{18.00} = .0500$	$A = 5.40$				
Then						
$X'_1 = .0190X_2 + .0006X_3 + .0199X_4 + .0500X_5 + 5.40$						
$r_{1-2345}^2 = (.230)(.55) + (.087)(.53) + (.185)(.52) + (.431)(.64) = .54465$						
$r_{1-2345} = .738$ , $\sigma_{1-2345} = 2.09\sqrt{1 - (.738)^2} = 1.40$						

As already specified, the correlations are written down in an order corresponding to equations (73) except that values to the left and below the diagonal are omitted. The first thing we do is to set up a check column. The first entry, 1.92, is obtained by summing, algebraically, the first row of correlations (including the diagonal 1.00); the second figure, 2.12, is the sum of the second row plus .56; the third entry, 1.99, is the sum of the third row plus .49 and .63; and the 1.63 is the sum of the fourth row plus .42, .46, and .39. The rule being followed should now be obvious: the  $j$ th entry in the check column is obtained by summing the 1.00 in the  $j$ th row with the values above it and to its right. The student should satisfy himself that this is equivalent to summing the correlations for the respective equations in (73). Since the check column will provide, at intervals, an automatic check on our computations, this summing should be done at least twice to insure accuracy.

Line (1) of the solution is obtained by copying down line (a), the first row of  $r$ 's; and line (2) consists of the line (1) values with the signs changed. The second part of the solution begins with line (3), which is obtained by copying down the (b) row of correlations. Line (4) is obtained by multiplying entries in line (1) by  $-.56$ , which figure is found in line (2) directly above the 1.000 of line (3). As indicated at the left, line (5) results from summing lines (3) and (4), i.e.,  $1.000 + (-.314)$  equals .686, etc.

At this point we have our first automatic check: summing line (5) across should yield 1.045, already obtained by vertical summing of values in the check column. To be a satisfactory check, these two sums should agree within limits consistent with errors imposed by rounding off to three decimal places. Acceptable discrepancies will be of the order  $\pm .001$ ,  $\pm .002$ ,  $\dots \pm .005$ , seldom larger.

Line (6) is obtained by multiplying line (5) by the negative reciprocal of its first entry. The correctness of the reciprocal used is evidenced by the fact that, when multiplied by .686, unity results. The ck attached to  $-1.524$  indicates that summing the entries in line (6) yields the same value as 1.045 multiplied by the negative reciprocal of .686, thus providing a further check. This completes the second part of the solution.

The third part begins with a copying of row (c) of the correlation table. The student should now be able to follow the steps; in

particular, he should note that a multiplier is secured from the last line of each preceding part of the solution; that each multiplier is applied in turn to the values in the line just above it; that, when all such multipliers have been utilized, the lines are summed (summing across again provides a check), and the resulting line is, as before, multiplied by the negative reciprocal of its first entry, thus completing the third part of the solution.

The fourth part involves similar operations. If we had five independent variables, we would proceed in like fashion, with an additional or fifth part. The schema can be extended to any number of variables. There will be as many parts to the solution as there are independent variables. The last part always consists of three columns of figures, and the bottom figure in the middle column is the value for  $\beta_n$ . In our example  $\beta_n = \beta_5 = .431$ .

The other  $\beta$ 's are determined by a "back" solution, which always involves a substitution of the value or values already found into the last line of the various parts (lines 11, 6, and 2 in our illustration). This back solution is given in Table 16. As a final check on all the computations, the four  $\beta$ 's obtained must be substituted into the four simultaneous equations with which we began. This check appears next in Table 16.

In order to put our results into useful form, we ordinarily require the multiple regression equation in raw score form, and for this we need the  $B$  coefficients and  $A$  as called for in formula (66) extended for more variables. To get the multiple correlation coefficient, the  $\beta$ 's and appropriate  $r$ 's are substituted in formula (69a), and from (70) we obtain the standard error appropriate for judging the accuracy of predictions made by the calculated regression equation. Table 16 includes these additional values.

If the problem involves analysis rather than prediction, one need not set up the regression equation or calculate the error of estimate. Appropriate interpretations would depend upon the  $\beta$ 's and  $r_{1.2345}$  (see discussion, pp. 176-177).

### SAMPLING ERRORS

The classical formula for the standard error of a multiple correlation involving  $n$  variables is

$$\sigma_{r_{1.23 \dots n}} = \frac{1 - r_{1.23 \dots n}^2}{\sqrt{N}} \quad (74)$$



If  $N$  is very large, say over 500, and if the value of  $r_{1.23 \dots n}$  is not too high, this formula will provide a satisfactory approximation. But when  $N$  is small and the number of variables,  $n$ , is large relative to the size of the sample, the above formula yields an underestimate of the error. The significance of the multiple correlation coefficient can best be ascertained by the analysis of variance technique, to be discussed in a later chapter.

Closely related to sampling is the shrinkage of the multiple correlation coefficient. This may be best understood by taking an extreme case. For the ordinary bivariate correlation, it is evident on a moment's reflection that, if  $N = 2$ , the correlation between the two variables must be perfect positive or perfect negative (it would be indeterminate if for either variable the two scores were the same); the regression line will pass through both plotted points on the scatter diagram. That is, in so far as prediction is concerned, there would be no error. In the case of three variables and  $N = 3$ , it would be possible to pass a plane through all three plotted points. In general, if  $n = N$ , we would get a perfect multiple  $r$ . Obviously  $N$  must be greater than  $n$  before any meaning can be attached to a multiple  $r$ . As  $n$  approaches  $N$ , the value of multiple  $r$  always approaches unity.

This suggests that, when  $n$  is large relative to  $N$ , the real significance of an obtained multiple  $r$  is questionable. In other words, the multiple correlation coefficient is subject to a positive bias, the magnitude of which depends upon the degree to which  $n$  approaches  $N$ . An unbiased estimate,  $r'$ , of the universe value of  $r_{1.23 \dots n}$  can be obtained from

$$r'_{1.23 \dots n} = \sqrt{1 - (1 - r^2_{1.23 \dots n}) \left( \frac{N - 1}{N - n} \right)} \quad (75)$$

This is sometimes known as a correction for shrinkage, since it has been observed that in general the correlation between observed and predicted values for a *new* sample tends to be less than the multiple  $r$  obtained by means of the  $\beta$ 's computed from the original sample. Obviously, if  $N$  is very large, say 500, and  $n$  small, say 10, the amount of bias or expected shrinkage is so small as to be negligible.

#### CAUTIONS AND REMARKS

As already indicated, there are two principal uses for the multiple correlation technique: (1) it yields the optimum weighting for

combining a series of variables in predicting a criterion and provides an indication of the accuracy of subsequent predictions; (2) it permits the analyzing of variation into component parts. There are certain more or less obvious pits into which the unwary user of the multiple regression and correlation method may fall. For example, it is possible to write a multiple regression equation for predicting school achievement ( $X_1$ ) from a knowledge of age ( $X_2$ ) and mental age ( $X_3$ ). In standard score form it might be  $z'_1 = .27z_2 + .67z_3$ , from which one might infer that school achievement depends upon age to a certain extent but upon mental age to a greater extent. However, it is entirely possible to argue that mental age depends partly upon school achievement. One could also use the same data to write the regression for age on mental age and school achievement; thus  $z'_2 = .56z_1 + .06z_3$ , from which the unwary might conclude that age depends upon school achievement and mental age.

Multiple correlation may be particularly deceptive when one has available several variables, each of which yields a rather low correlation with the criterion and from which those yielding the higher correlations with the criterion are selected for the prediction equation. Such selecting tends to capitalize on correlations which might be high because of sampling fluctuations. For example, the author was once requested to compute the multiple  $r$  for an 11-variable problem. None of the 10 variables showed a very high correlation with the criterion, the highest being .27. The resulting multiple was .44, which was statistically significant for the sample of 89 cases. When it was learned that 10 variables out of 40 had been selected as the most promising, i.e., because they showed the highest correlations with the criterion, the real significance of the multiple  $r$  of .44 was questioned. That it really was misleading was clearly evidenced by the fact that for a second and similar sample the variable originally yielding the highest  $r$  (.27) now produced an  $r$  of  $-.11$ . That is, the supposedly best single predictor was actually of very doubtful value, and this, coupled with a tendency for the next highest  $r$ 's to drop appreciably, meant that predictions by the regression equation could not be as good as was inferred from the multiple of .44.

Nothing has been said as yet concerning the principal assumption and consequent limitation in the use of multiple regression equations, namely, that regressions for the first-order correlations

must be linear. There are methods for handling multiple correlation for curvilinear regressions. The reader is referred to M. Ezekiel's *Methods of correlation analysis*.<sup>†</sup>

It is not obvious from our discussion that, in general, the increase in the multiple correlation which results from adding variables beyond the first five or six is very small. This phenomenon of diminishing returns would not, of course, operate if we were to find an additional variable which correlated much more highly with the criterion than any of those already utilized.

Another fact which may not be apparent to the reader is that we can expect the multiple  $r$  to be higher when the intercorrelations among the predictors are low instead of high. This point can be easily demonstrated to one's own satisfaction by computing the multiples for, say,  $r_{12} = .50$ ,  $r_{13} = .50$ , and varying values for  $r_{23}$ .

An interesting paradox of multiple correlation and an exception to the fact mentioned in the previous paragraph is that it is possible to increase prediction by utilizing a variable which shows no, or low, correlation with the criterion, *provided* it correlates well with a variable which does correlate with the criterion. Thus, if  $r_{12} = .400$ ,  $r_{13} = .000$ , and  $r_{23} = .707$ , the regression equation will be  $z'_1 = .800z_2 - .566z_3$ , and  $r_{1 \cdot 23}$  will equal .566. It is thus seen that, when  $z_3$  is combined with  $z_2$ , an appreciable gain in prediction occurs even though when taken alone  $z_3$  is worthless as a predictor of  $z_1$ .

Such a variable has been termed a "suppressant." One does not quickly see just how a suppressant variable, showing no correlation with the criterion, can increase the accuracy of prediction. Perhaps this point can be explained by reasoning by way of the notion that correlation can be thought of in terms of common elements (pp. 140-141). Suppose that  $X_1$  is composed of 10 elements,  $X_2$  of 10,  $X_3$  of 5, and suppose that  $X_1$  and  $X_2$  have 4 elements in common,  $X_2$  and  $X_3$  have 5 elements in common, and  $X_1$  and  $X_3$  have no overlapping elements. Diagrammatically, the variables and elements would be

$$\begin{array}{cccccccccc} & & & & X_1 & & & & X_3 & \\ & & & & \hline a & a & a & a & a & b & b & b & b & c & d & d & d & d & d \\ & & & & & & & & & & X_2 & & & & \hline \end{array}$$

<sup>†</sup> Ezekiel, M., *Methods of correlation analysis*, New York: John Wiley, 1941

By substituting in the common element formula for correlation, we find  $r_{12} = .400$ ,  $r_{13} = .000$ ,  $r_{23} = .707$ . These lead to  $z'_1 = .800z_2 - .566z_3$ , and  $r_{1\cdot23} = .566$ . Variable  $X_3$  has a negative regression weight, i.e., by the use of  $X_3$  something is being subtracted or suppressed. As set up here for illustrative purposes, all the elements of  $X_3$  are contained in  $X_2$ ; these elements are not related to  $X_1$  and hence their presence in  $X_2$  must tend to lower the correlation between  $X_1$  and  $X_2$ ; if these elements could be suppressed, the correlation between  $X_1$  and  $X_2$  minus the irrelevant (so far as  $X_1$  is concerned) elements of  $X_2$  should be higher than  $r_{12}$ . Actually, if we think of the "d" elements of the diagram as being nonexistent, we would have variation in  $X_2$  dependent upon only 5 elements, 4 of which overlap with  $X_1$ . The correlation between  $X_1$  and the abridged  $X_2$  would be  $4/\sqrt{10(5)}$  or .566, which has exactly the same value as the multiple  $r$  obtained above. This exact correspondence to  $r_{1\cdot23}$  will be obtained only when all the  $X_3$  elements are contained in  $X_2$ . If  $X_3$  contains other elements, its use as a suppressant will aid in predicting  $X_1$ , but the resulting  $r_{1\cdot23}$  will not correspond to an  $r$  deducible from the common element formula. The reason for this is left as an exercise.

The student, by resort to the notion of common elements, may secure a better understanding of the proposition that a higher multiple is obtainable when the correlations with the criterion are high and the correlations between the predictors low or zero. The reader should be warned, however, that such a condition is hard to realize in practice, as is also the finding of variables which will qualify as suppressants.

# NOTE ON NOTATION

The symbol  $r_{1\cdot23}$  has been used to represent the correlation (multiple) between  $X_1$  and the best combination of  $X_2$  and  $X_3$ . This should not be confused with  $r_{12\cdot3}$ , which indicates the correlation (partial) between  $X_1$  and  $X_2$  with the effect of  $X_3$  ruled out or held constant. The symbol  $\sigma_{y\cdot x}$ , it will be recalled, stood for the standard error of estimate of  $Y$  as estimated from  $X$ ;  $\sigma_{1\cdot2}$  would be the error of  $X_1$  when estimated from  $X_2$ ; and  $\sigma_{1\cdot23}$  would be the standard error of estimate of  $X_1$  when estimated from  $X_2$  and  $X_3$  by means of the multiple regression equation.

In the foregoing discussion,  $\beta_2$  has been used as the symbol for the regression weight of  $X_2$ . A more formal, albeit cumbersome, notation would be  $\beta_{12 \cdot 345}$ , which would be read as the regression of  $X_1$  on  $X_2$ , i.e., the coefficient for  $X_2$ , when used in combination with  $X_3$ ,  $X_4$ , and  $X_5$ . It is not an accident that the subscript pattern resembles that for the partial correlation coefficient. If we were dealing with a three-variable problem,  $\beta_2$  could be written as  $\beta_{12 \cdot 3}$ . This notation really means that we have the net regression of  $X_1$  on  $X_2$  when  $X_3$  is held constant. Hence the coefficients are sometimes spoken of as *partial* regression coefficients. As a matter of fact, these partial or multiple regression coefficients can be computed by way of partial correlation coefficients, but the method is not nearly so straightforward and self-checking as the Doolittle procedure.

## Other Correlation Methods

The product moment correlation measure is applicable only when the two variables are graduated, is restricted by the assumption of linearity of regression, and needs careful qualifying if either or both variables yield skewed distributions. There are, therefore, many problems for which it is inappropriate. In general, the majority of the situations which are met in practice can be handled by some type of correlational technique.

There are no general rules to follow in the case of variables yielding skewed distributions. Frequently, one can use a logarithmic transformation of such a variable and thereby secure scores which are at least approximately normal; or one may deliberately normalize the distribution by converting the raw scores into *T* scores. When one considers the arbitrary units involved in most psychological measurement, such a procedure would seem not only permissible but also defensible in that the correlational description of the relationship need not be qualified because of skewness.

The situations arising most frequently in practice, for which measures of correlation are apt to be needed, can be subsumed under the following six headings: (1) graduated measures for one variable, dichotomized or two-category information for the second variable; (2) both variables dichotomized; (3) three or more categories for one variable and two or more for the second; (4) three or more categories for one variable and a graduated series of measures for the other; (5) both variables graduated, with curvilinear relationship; (6) when data are rank-orders.

An estimate of the degree of correlation for each of the above situations can be obtained provided certain assumptions concerning the variables can be regarded as tenable. Ordinarily the



graduated variable can be thought of either as being continuous or as progressing in a sufficient number of discrete steps so as to give the appearance of continuity. The approach to normality for such series can, obviously, be specified. The nature of the categorized variable, whether discrete or continuous, can ordinarily be ascertained on logical grounds, but the question of whether a continuous variable for which we have only a distribution by categories would yield a normal distribution if we had some measuring stick for the trait is not easy to answer.

### BISERIAL CORRELATION

When one variable is measured in a graduated fashion and the other is in the form of a dichotomy, we have the so-called biserial situation, for which there are 2 measures of correlation: biserial  $r$  and point biserial  $r$ . The difference between these 2 measures depends essentially on the type of assumption which is made concerning the nature of the dichotomized variable.

Table 17. BISERIAL TABLE FOR "ABSTRACT WORDS" AS  $X$  AND BINET IQ AS  $Y$

IQ	Item		Totals	
	Fail (1)	Pass (2)		
145-149		1	1	
140-145				$\sigma_y = 17.69$
135-139		1	1	$p_1 = .37$
130-134		3	3	$p_2 = .63$
125-129		4	4	$z = .378$
120-124		6	6	
115-119		10	10	
110-114		7	7	$r_b = \frac{(109.86 - 84.43)(.37)(.63)}{(.378)(17.69)}$
105-109	1	8	9	
100-104	1	5	6	$= .89$
95-99	4	9	13	Or by formula (76a):
90-94	7	6	13	
85-89	9	2	11	$r_b = \frac{(109.86 - 100.45)(.63)}{(.378)(17.69)}$
80-84	3	1	4	
75-79	4		4	$= .89$
70-74	5		5	
65-69				$r_{pb} = \frac{(109.86 - 100.45)}{17.69} \sqrt{\frac{.63}{.37}}$
60-64	3		3	
Totals	37	63	100	$= .69$
Means	84.43	109.86	100.45	

The most typical example of situations calling for one or the other of these measures is to be found in the test (mental and personality) field: the correlation between an item scored as pass or fail (yes or no, like or dislike, etc.) and a graduated criterion variable (or a total score on all of a set of items). We need to know each individual's score on the graduated variable and the dichotomy to which he belongs. Then we can make a distribution or scattergram with from 12 to 20 intervals for the graduated variable along the  $y$  axis, and with 2 intervals for the 2 categories along the  $x$  axis. Such a correlation scattergram is given in Table 17, which involves pass-fail on "abstract words" vs. composite IQ on Forms L and M of the 1937 Stanford-Binet. It is obvious that there is a tendency for those who fail the item to have lower IQ's than those who pass—performance on the item is related to IQ.

**Biserial coefficient,  $r_b$ .** If it can be assumed that underlying the dichotomy there is a continuous variable, we can obtain a measure of correlation which is an estimate of what the product moment correlation would be in case the dichotomous variable were measured in such a way as to produce a normal distribution. This estimate is given by

$$r_b = \frac{(M_2 - M_1)(p_1 p_2)}{z \sigma_y} \quad (76)$$

or by the exact equivalent

$$r_b = \frac{(M_2 - M_y)p_2}{z \sigma_y} \quad (76a)$$

in which  $p_1$  = proportion of cases in the first category.

$p_2$  = " " " " " second " "

$M_1$  = mean of  $Y$ 's for cases in the first category.

$M_2$  = " " " " " second "

$M_y$  = " " all the  $Y$  scores.

$\sigma_y$  =  $SD$  of " " " " "

$z$  = ordinate for the unit normal curve at the point where  $p_1$  (or  $p_2$ ) cases are cut off; it is determined by entering  $p_1$  or  $p_2$ , whichever is smaller, as a  $q$  value in Table A, then reading off the adjacent ordinate value in the fourth column of the table (interpolating if necessary).

Formula (76a) is the more convenient when each of a series of items is to be correlated against the same graduated variable. The computations are illustrated in Table 17.

In the derivation of  $r_b$  it is assumed not only that a normal distribution underlies the dichotomy but also that the regressions would be linear if the dichotomized variable were measured. The latter assumption cannot be checked; it is apt to hold for ability variables but may be violated for personality traits. The former assumption has troubled many. Actually, the main issue is the question of continuity. Consider the pass-fail dichotomy; it is obvious that failing a test item represents anything from a dismal failure up to a near pass, whereas passing the item involves barely passing up to passing with the greatest of ease. Such a line of reasoning is certainly presumptive evidence for continuity, and a similar argument can be advanced as regards yes-no, like-dislike, and similar categories. Given a continuous trait, it is usually (if not always) possible to construct a test thereof which yields a normal distribution, and consequently we need not worry about the mathematical assumption of normality when using  $r_b$ . We can justify the use of  $r_b$  with obviously continuous variables by saying, as pointed out earlier, that the obtained coefficient represents what we would expect the product moment correlation to be if we had a measuring scale, for the dichotomized trait, which yielded a normal distribution.

The sampling error of biserial  $r$  is given approximately by

$$\sigma_{r_b} = \frac{\frac{\sqrt{p_1 p_2}}{z} - r_b^2}{\sqrt{N}} \quad (77)$$

As an exercise, the student should compare the magnitude of the sampling error of biserial  $r$  for various cuts ( $p$  values) with that of the product moment  $r$  as given by the analogous classical form,  $\sigma_r = (1 - r^2)/\sqrt{N}$ . It might be anticipated that the sampling error will be large when dichotomies are extreme, i.e., involve cuts yielding very high (and low)  $p$ 's. Thus, if  $N = 100$ , and we have a .95-.05 cut, it follows that one of the means used in computing  $r_b$  by formula (76) will be based on only 5 cases and therefore will be subject to rather large sampling fluctuation, which incidentally

will not be counterbalanced entirely by the relatively greater stability of the other mean. It may occur to the reader that the use of formula (76a) would overcome this difficulty, since one can always arrange to use the mean for the category having the larger number of cases, thereby avoiding the unstable mean. This appears plausible enough; its refutation is left to the student.

The fact that the sampling error for biserial  $r$  is large when extreme dichotomies are involved should serve as a warning. Unless  $N$  is fairly large, one should not place much confidence in a biserial  $r$  based on cuts more extreme than .10 (or .90).

Since no  $r$  to  $z$  transformation is available for use with biserial  $r$ , the difficulty of skewed sampling distributions for high  $r_b$ 's cannot be overcome. In testing the null hypothesis (that no correlation exists), the  $r$  term in formula (77) may be dropped. For  $N$  small, a more adequate test of the significance of  $r_b$  is possible by way of the  $t$  test for the difference between  $M_2$  and  $M_1$ .

Although  $r_b$  is an estimate of a product moment  $r$ , there are limitations as to its interpretation. It is, of course, a measure of the degree of relationship between 2 variables. It does not, however, enter into prediction formulas, nor does it lead to an error of estimate. If we know to which  $X$  category an individual belongs, the predicted  $Y$  is simply the mean of the  $Y$  scores for that category, and the error of such an estimate is the standard deviation of the  $Y$  scores in the given category. This error of estimate would not equal  $\sigma_y \sqrt{1 - r_b^2}$ .

If we have a  $Y$  score to use in predicting an individual's  $X$  category, we estimate on the basis of the tendency for those possessing  $Y$  scores in a given interval to fall predominantly into the first or second category on  $X$ . The error for such a prediction must depend upon the relative frequencies in these 2 categories for individuals possessing a given  $Y$  score. Thus, if the frequencies in the first and second categories were 18 and 6 (for a given  $Y$  interval), the error might be stated something like this: the odds are 3 to 1 that the given individual's  $X$  position is in the first category; i.e., 75 per cent of the time the prediction would be correct. But such a percentage statement might itself be subject to grave sampling error since it is based on a small  $N$ ; and such a statement of error might need to be qualified according to the  $p$ 's. Why?

**Point biserial,  $r_{pb}$ .** If the dichotomous trait is truly discrete, an appropriate measure of correlation is given by

$$r_{pb} = \frac{(M_2 - M_1)\sqrt{p_1p_2}}{\sigma_y} \quad (78)$$

or its equivalent

$$r_{pb} = \frac{M_2 - M_1}{\sigma_y} \sqrt{\frac{p_2}{p_1}} \quad (78a)$$

Actually,  $r_{pb}$  is the product moment correlation between  $Y$  and the  $X$  categories scored as either 0 or 1 (scoring as 1 and 2, or as 4 and 10, or any other 2 values will yield the same correlation). The value of  $r_{pb}$  for the data of Table 17 is .69, compared to an  $r_b$  of .89. The magnitude of  $r_{pb}$  tends to be less than that of  $r_b$  for the same set of data, as can be seen by examining the following connection between the 2 coefficients:

$$r_{pb} = \frac{z}{\sqrt{p_1p_2}} (r_b)$$

For a 50-50 dichotomy,  $z = .3989$  and  $r_{pb} = .798r_b$ , and as the dichotomy departs farther and farther from 50-50 the discrepancy between  $r_{pb}$  and  $r_b$  increases. For a 10-90 cut we have  $r_{pb} = .585r_b$ . The maximum degree of correlation between a dichotomous variable and a normally distributed variable will occur when there is no overlap between the  $Y$  distributions for the 2 categories. For such a situation  $r_b$  will be either +1.00 or -1.00 regardless of the cut, whereas  $r_{pb}$  will be  $\pm .798$  for a 50-50 cut and only  $\pm .585$  for a 10-90 cut. These 2 coefficients are not on the same scale; they will agree only when there is exactly no relationship between the 2 variables. Even if the dichotomous variable were a genuine point variable,  $r_{pb}$  as an expression of the degree of relationship would not be comparable either to  $r_b$  or to the product moment  $r$  between 2 variables measured in a graduated fashion.

Despite the fact that true point variables are practically nonexistent in psychology and despite the difficulties of interpreting  $r_{pb}$  as a terminal descriptive statistic,  $r_{pb}$  has a rightful place in

certain analytical and practical work where the 2 categories are arbitrarily, for convenience, assigned point scoring values of, say, 0 and 1. For example, if a dichotomized variable with point scoring were included in an  $n$  variable multiple regression equation, point biserial  $r$ 's would be the correct values for the correlation of the dichotomized variable with the remaining  $n-1$  variables.

For the large sample situation the significance of  $r_{pb}$  (as a deviation from zero) may be tested by using  $\sigma_{r_{pb}} = 1/\sqrt{N}$  as its standard error. For small samples, the  $t$  test for the difference,  $M_2 - M_1$ , is appropriate.

A troublesome difficulty with the biserial coefficient,  $r_b$ , is that it occasionally exceeds unity. The usually given explanation for this is that the assumption of normality for the dichotomous variable is not tenable, but it seems more likely that when such  $r$ 's occur it is because the graduated variable, for the combined categories, is either platykurtic or bimodal in distribution.

### TETRACHORIC CORRELATION

When both variables yield only dichotomized information, as, for example, 2 items scored as passed or failed, it is possible to secure an estimate of what the correlation would be if the underlying traits were continuous and normally distributed or if they were so measured as to give normal distributions. The measure of relationship for such a situation is known as the *tetrachoric correlation coefficient*, usually designated as  $r_t$ . It is not feasible to derive here the formula for tetrachoric correlation, but perhaps a few words will help one understand the reasoning back of the formula.

Let us suppose that we have before us a scattergram for the correlation between height and weight; let us further assume that this scatter exhibits all the characteristics of a normal correlational surface as defined by equation (41). That is, the 2 marginal distributions and all the vertical and horizontal array distributions are normal; the regressions are linear; and the arrays homoscedastic. For such a normal plot, it is possible, knowing the degree of correlation and the means and sigmas of the 2 variables, to specify how many or what proportion of the cases will fall in any given segment of the scatter plot. This can be done by mathematical manipulation of formula (41) or by the aid of



Table VIII of Pearson's *Tables for statisticians and biometricians, part II*.\*

Now, of course, if one had placed before him a scatter for height vs. weight and were asked how many cases fell in that portion of the table below 120 pounds *and* also below 68 inches, he would simply count them. But suppose he were told that, when the 2 axes were cut at 120 pounds and 68 inches, the frequencies in each of the 4 quadrants so formed were as shown in Table 18.

Table 18. CORRELATION FOR HEIGHT AND WEIGHT DICHOTOMIZED

	Below 120 lb.	Above 120 lb.	
Above 68 in.	10	80	90
Below 68 in.	60	50	110
	70	130	200

The purpose of tetrachoric correlation is to ascertain the degree of correlation which would permit the observed frequencies in such a fourfold table. A more rigorous statement would be: (Given the 4 frequencies, what should be the true correlation—for the scatter underlying the fourfold table—in order to make the obtained 4 frequencies most likely?)

In order to secure this estimate it is necessary to convert into a proportion each of the 4 frequencies and each of the marginal totals by dividing by  $N$ . For the fourfold table we may symbolize the frequencies as in Table 19, the proportions as in Table 20.

Table 19. FREQUENCIES

	-	+	
+	A	B	A + B
-	C	D	C + D
	A + C	B + D	N

Table 20. PROPORTIONS

	-	+	
+	a	b	p
-	c	d	q
	q'	p'	1.0

Then, the tetrachoric coefficient can be obtained from the following rather forbidding equation:

\* Pearson, Karl, *Tables for statisticians and biometricians, part II*, Cambridge: Cambridge University Press, 1931.

$$\frac{c - qq'}{z_x z_y} = r + xy \frac{r^2}{2} + (x^2 - 1)(y^2 - 1) \frac{r^3}{6} + (x^3 - 3x)(y^3 - 3y) \frac{r^4}{24} + \dots \quad (79)$$

in which it is assumed that both  $q$  and  $q'$  are less than .50. The general rule is to choose whichever is smaller,  $p$  or  $q$ , to pair with whichever is smaller,  $p'$  or  $q'$ . This determines, logically, whether  $a$  or  $b$  or  $c$  or  $d$  becomes a part of the formula. Thus one can have  $c - qq'$  (as given), or  $b - pp'$ , each of which will yield a positive  $r$  for positive correlation or a negative  $r$  for negative correlation, or one can have  $a - q'p$  or  $d - qp'$ , each of which will yield an  $r$  with sign opposite to its true sign. (It is, of course, here assumed that reading to the right on the  $x$  axis and up on the  $y$  axis means *more* of the traits.)

We must next specify the meaning of the  $x$ ,  $y$ , and  $z$ 's in formula (79). As for biserial  $r$ ,  $z_y$  is the ordinate of the unit normal curve where  $q$  proportion of the cases are cut off;  $z_x$  has a similar meaning for  $q'$ . The  $y$  represents the value on the base line of the unit normal curve where  $q$  cases are cut off, i.e., the  $x/\sigma$  in Table A of the Appendix, and  $x$  is similarly determined from a knowledge of  $q'$ .

To equation (79) additional terms may be added which will result in a closer approximation at the expense of a greater, if not an impossible, amount of computation. For the given formula, the solution for  $r$  involves determining the roots of a fourth-degree or quartic equation. Either Horner's or Newton's methods, as described in college algebra texts, will do the trick. The fourth-degree equation will yield satisfactory approximations except when  $r$  is high.

The solution of a quartic equation is not difficult, nor is it so easy as to lead to mass production of tetrachoric  $r$ 's. Fortunately, it is no longer necessary to go through this tedious method for getting an approximation to the value of  $r$ . Diagrams † are available which enable one to determine quickly the value of  $r$  for any given table of proportionate frequencies. Anyone having as many as a half-dozen tetrachorics to compute will find it economical to possess a copy of these diagrams.

† Chesire, L., Saffir, M., and Thurstone, L. L., *Computing diagrams for the tetrachoric correlation coefficient*, Chicago: University of Chicago Bookstore, 1933.

The tetrachoric  $r$  is particularly useful in estimating the degree of correlation between variables for which we have only dichotomized information, but it can also be used instead of biserial  $r$  or the product moment  $r$ , since situations for which these 2 methods apply can readily be converted into fourfold tables by simply dichotomizing the graduated variables. The advantage of so estimating correlation is that tetrachoric  $r$  is much easier to determine (by using the computing diagrams) than is calculating either biserial  $r$  or the product moment  $r$ . Indeed, this fact of computational economy has led a number of investigators to use  $r_t$  when product moment  $r$ 's could be determined. That such a practice may be short-sighted economy becomes quite evident when we turn to the sampling fluctuation of  $r_t$ .

The standard error of  $r_t$  is closely approximated by

$$\sigma_{r_t} = \frac{\sqrt{pq p' q'}}{z_x z_y \sqrt{N}} \sqrt{(1 - r^2) \left[ 1 - \left( \frac{\sin^{-1} r}{90^\circ} \right)^2 \right]} \quad (80)$$

When this is compared to the classical formula for the standard error of a product moment  $r$ , i.e., to  $\sigma_r = (1 - r^2)/\sqrt{N}$ , it will be seen that the tetrachoric  $r$  has a much larger sampling error. To illustrate the difference, the sigmas for 4  $r$ 's for 2 different dichotomies are presented in Table 21 along with the sigmas (by the classical formula) of the corresponding product moment  $r$ 's for  $N = 100$ .

Table 21 SAMPLING ERRORS OF  $r_t$  AND  $r$  COMPARED

$r$ or $r_t$	$p$	$p'$	$\sigma_{r_t}$	$\sigma_r$
.00	.50	.50	.157	.100
.00	.80	.80	.204	.100
.40	.50	.50	.130	.084
.40	.80	.80	.182	.084
.60	.50	.50	.115	.064
.60	.80	.80	.150	.064
.80	.50	.50	.073	.036
.80	.80	.80	.095	.036

It can readily be seen from this table that  $r_t$  is much less stable than  $r$ ; in fact, even for the most favorable comparison (.50-.50

cuts, low  $r$ 's), the standard error of the tetrachoric coefficient is more than 50 per cent greater than that for the product moment coefficient. This means that one must have more than twice as many cases to attain the same degree of sampling stability for a tetrachoric as for a product moment correlation coefficient. For .80 .20 cuts and low correlations, 4 times as many cases are needed to have comparable sampling errors. For high correlations and also for more extreme cuts,  $r_t$  compares still less favorably with  $r$ .

The foregoing discussion and further study of formula (80) lead to 2 obvious conclusions.

First, the increasing sampling instability of  $r_t$  as the dichotomies become more extreme warns us that, unless  $N$  is large, one cannot place much reliance on  $r_t$  for cuts more extreme than .10 .90; seldom will  $N$  be large enough to warrant confidence in a tetrachoric based on cuts more extreme than .05 .95.

Second, in using  $r_t$  instead of the product moment  $r$  when the latter is calculable, one is always throwing away the equivalent of more than half the available data. Thus the computational economy may be an expensive luxury—it is very doubtful whether the calculation of a product moment  $r$  for  $N$  cases will ever require anything but a fraction of the expense of securing data on the additional  $N$  cases needed to counterbalance the greater sampling error incurred in using the tetrachoric coefficient.

As in the case of  $r_b$ , no  $r$  to  $z$  transformation exists for handling the sampling errors of high tetrachorics. For testing the null hypothesis, that  $r_t$  for the universe is zero, we may use a simpler expression for its standard error, namely,  $\sigma_{r_t} = \sqrt{pq p' q'} z_x z_y \sqrt{N}$ . Another method for judging the significance of the correlation computed from a fourfold table will be presented in the next chapter.

The use of tetrachoric  $r$  is circumscribed by an assumption that the underlying correlational surface is of the normal type. Among other things this implies (1) that the dichotomized traits are continuous and normally distributed, and (2) that the regressions are linear. Although, as discussed in connection with biserial  $r$ , we are usually ignorant of the tenability of (1), this ignorance can be partially overcome by regarding the correlation as that which would obtain if the traits were normalized; i.e., it can be argued that the use of tetrachoric  $r$  automatically normalizes the distribu-

tions. It is not so easy to dispose of assumption (2), since the normalizing of variables will not necessarily lead to linearity of regression. The only consolation here is that *measured* psychological traits are usually linearly related, if related at all.

#### FOURFOLD POINT CORRELATION

If we can safely assume point distributions for both dichotomous variables, a descriptive measure of correlation can be obtained from a fourfold table (Table 19) by

$$r_p = \frac{BC - AD}{\sqrt{(A + B)(C + D)(A + C)(B + D)}} \quad (81)$$

or from the table of proportionate frequencies (Table 20) by the exact equivalent

$$r_p = \frac{c - qq'}{\sqrt{pqp'q'}} \quad (81a)$$

The fourfold point correlation coefficient is frequently referred to as the *phi* coefficient and designated by  $\phi$ . Actually, it is the product moment correlation between the 2 variables each scored in a point fashion (say, 0 and 1). Unlike the point biserial,  $r_p$  can be unity but only when  $p = p'$ . Otherwise (i.e., in nearly all situations)  $r_p$  and  $r_t$  from the same table will differ in value, with  $r_p$  being lower, and the difference between the 2 becomes greater as the dichotomy for either variable, or both, varies farther and farther from 50-50.

A few examples will illustrate the difference in the magnitude of  $r_p$  and  $r_t$ . It is possible to have a fourfold table with 50-50 and 50-50 cuts which yields an  $r_t$  of .50 and an  $r_p$  of .32, and a table with 16-84 and 16-84 cuts which yields an  $r_t$  of .50 and an  $r_p$  of only .26. For similar tables (as regards cuts) we may have  $r_p$  values of .59 and .52 when  $r_t$  is .80. Thus,  $r_p$  is not interpretable on the same scale as  $r_t$  (or  $r$  or  $r_b$ ) as a measure (terminal descriptive statistic) of the degree of relationship.

However,  $r_p$  is useful (and necessary) in certain analytical work. If variable  $U$  and variable  $V$  were dichotomous and each scored as 0 and 1, then  $r_p$  would be the appropriate value to use in formula (37), p. 137, to obtain the variance of  $W$ , defined as  $U + V$ . If formula (20), p. 59, for the standard error of the difference

between correlated proportions were written analogously to formula (25b), p. 85,  $r_p$  would be used. It is also used in the statistical theory of mental tests.

For testing whether  $r_p$  deviates significantly from zero we may safely use  $1/\sqrt{N}$  as its standard error when  $N$  is not small.

### CONTINGENCY COEFFICIENT

The *contingency coefficient* is a measure of the degree of association or correlation which exists between variables for which we have only categorical information. The number of categories can be such as to provide a 2 by 2 table (as for tetrachoric correlation) or a 2 by 3, or a 3 by 3, or a 3 by 4, or a 4 by 4, or a  $k$  by  $l$  table. This coefficient is stated in terms of a quantity known as  $\chi^2$  (*chi square*) thus

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}} \quad (82)$$

where

$$\chi^2 = \sum \frac{(O - E)^2}{E} \quad (83)$$

in which  $O$  is the observed frequency (not percentage) and  $E$  is the expected frequency for a given cell. In a 2 by 3 table there would be 6 cells, hence 6 values summed to get  $\chi^2$ . The expected cell frequencies for the contingency situation are those frequencies which would exist if there were no association or relationship between the given variables. It can thus be anticipated that, the larger the discrepancy between expected and observed frequencies *relative* to the expected, the larger the value of  $\chi^2$  and consequently the higher the value of  $C$ .

An example will help to clarify the above. Suppose that we have 2 variables, each of which yields 3 categories or classifications, and that the observed frequencies are as given in Table 22, which also contains the expected frequencies in parentheses. (Fictitious data; marginal frequencies arranged so as to simplify exposition.) In order to ascertain the expected frequencies needed in the computation of  $\chi^2$ , we ask what cell frequencies would be expected if there were no relationship, or zero association, between the 2 variables. Consider the 100 classified as college; if no association



Table 22. CONTINGENCY TABLE

	Low	Medium	High	
College	5 (20)	45 (60)	50 (20)	100
High school	50 (40)	110 (120)	40 (40)	200
Grade school	45 (40)	145 (120)	10 (40)	200
	100	300	100	500

existed, one would expect that these 100 would be distributed according to a 1, 3, 1 ratio, i.e., in the same ratio as the marginal frequencies at the bottom. Thus the expected cell frequencies for the top row of cells would be 20, 60, 20. The expected frequencies for the middle and bottom rows of cells should also be in a 1, 3, 1 ratio. Both these rows would have expected frequencies of 40, 120, 40.

It will be noted that (1) the expected frequencies for the *columns* follow, as they should, the ratio of 1, 2, 2, i.e., the ratio of 100, 200, 200 for the marginal frequencies on the right; (2) the expected frequencies sum to the same marginal totals as the observed frequencies; and (3) the expected frequencies actually exhibit a zero relationship between the 2 characteristics.

In practice, the computation of the expected frequencies can readily be accomplished by either of 2 schemes: (1) express the marginal totals along the bottom as proportions of the total  $N$ , then multiply each of the frequencies on the right margin by each proportion in turn, entering the resulting product in the cell common to the 2 marginal figures involved in the multiplication; or (2) multiply any frequency on the bottom margin by any frequency on the right margin, and then divide this product by  $N$ ; the result is the expected frequency for the cell common to the 2 marginals involved in the products.

The computation of  $\chi^2$  is now a routine matter. We simply take each cell in turn, square the difference between the observed and expected value, and divide by the expected frequency. Thus we have

$$\begin{aligned}
 (5 - 20)^2/20 &= 11.25 \\
 (45 - 60)^2/60 &= 3.75 \\
 (50 - 20)^2/20 &= 45.00 \\
 (50 - 40)^2/40 &= 2.50 \\
 (110 - 120)^2/120 &= .83 \\
 (40 - 40)^2/40 &= .00 \\
 (45 - 40)^2/40 &= .62 \\
 (145 - 120)^2/120 &= 5.21 \\
 (10 - 40)^2/40 &= 22.50
 \end{aligned}$$

The sum of these quantities, 91.66, is  $\chi^2$ . To get  $C$ , the coefficient of contingency, the value of  $\chi^2$  is substituted in formula (82), thus

$$C = \sqrt{\frac{91.66}{500 + 91.66}} = .39$$

This strength of association is not to be interpreted as indicating the same degree of relationship as an ordinary (or biserial or tetrachoric) coefficient of the same magnitude. One reason for this is that the upper limit for the contingency coefficient is a function of the number of categories. The upper limit for a 2 by 2 table is  $\sqrt{\frac{1}{2}}$ ; for a 3 by 3 table,  $\sqrt{\frac{2}{3}}$ ; for a 4 by 4 table,  $\sqrt{\frac{3}{4}}$ ; for a 5 by 5 table,  $\sqrt{\frac{4}{5}}$ ; for a  $k$  by  $k$  table,  $\sqrt{(k-1)/k}$ . The exact upper limits for rectangular tables, such as 2 by 3, 2 by 4, 3 by 4, are unknown. (As an exercise, the student might demonstrate to his own satisfaction the upper limit for 2 by 2 and 3 by 3 tables.) The reader will also note that  $C$  can never be negative.

Despite having varying maximal values, contingency coefficients have a decided advantage over other measures of relationship; no assumptions involving the nature of the variables need be met—continuous or discrete variables, normal or skewed or any shaped distributions for underlying traits, ordered or unordered series, and combinations thereof are permissible.

Disadvantages are that any 2 contingency coefficients are not comparable unless derived from tables of the same size, that they are noncomparable to product moment  $r$ 's (and estimates thereof) unless certain corrections are applied, and that the formula for sampling error is unwieldy. The necessary corrections and the sampling error formula may be found in Kelley,<sup>†</sup> but before con-

<sup>†</sup> Kelley, T. L., *Statistical method*, pp. 266-271, New York: Macmillan, 1924.

sulting Kelley, the reader might bear in mind the following comments.

In regard to the corrections, the first is for number of categories. The additional correction to make  $C$  an estimate of  $r$  involves the assumption that the underlying traits are continuous and normal in distribution. Furthermore, this correction is very tedious to make. It is suggested that, if the assumption of normally distributed continuous variables is tenable, one is justified in reducing a contingency table of more than 4 cells to a 2 by 2 table and then determining the value of tetrachoric  $r$ . When reducing to a fourfold table, one should combine adjacent categories so as to have dichotomies as near to .50-.50 proportions as possible. The combination should *not* be made on the basis of the pattern of cell frequencies, since this is likely to involve a capitalization or decapitalization on chance. One might take several or all possible fourfold combinations, thus securing several tetrachoric  $r$ 's which may then be averaged.

As to the unwieldy sampling error formula for  $C$ , it is suggested that in so far as one wishes simply to test the null hypothesis, i.e., that there is no relationship between the 2 given variables, one need only enter the value of  $\chi^2$  into an appropriate probability table to test its significance. If  $\chi^2$  is significant, then the relationship is significantly greater than zero. This use of  $\chi^2$  will be discussed in the next chapter. It should be remarked that, if any one (or more) expected cell frequency is small, say less than 5, the resulting  $C$  may be quite erroneous.

Chi square for a fourfold table can be readily obtained by formula without first computing expected frequencies. Thus for a set of frequencies like that of Table 19 we have

$$\chi^2 = \frac{N(AD - BC)^2}{(A + B)(C + D)(A + C)(B + D)}$$

This resembles formula (81). In fact, there is a relationship between the fourfold point coefficient ( $r_p$ ),  $\chi^2$ , and  $C$ :

$$r_p^2 = \frac{\chi^2}{N} \quad \text{and} \quad C = \sqrt{\frac{r_p^2}{1 + r_p^2}}$$

Other measures of association or of correlation between attributes have been advocated. This is not the place to argue the pros

and cons of these other measures. It seems to the author that the measures we have discussed are the more defensible.

### THE CORRELATION RATIO OR $\eta$ (ETA)

It will be recalled that one way of understanding the product moment correlation coefficient is to note from the relationship,  $r^2 = 1 - \sigma_{y \cdot x}^2 / \sigma_y^2$  (or  $r^2 = 1 - \sigma_{x \cdot y}^2 / \sigma_x^2$ ), that the degree of correlation is a function of the error of estimate variance relative to the total variance of the variable being predicted by a linear regression line. If the array means fail to fall on a straight line, it can rightly be argued that better prediction can be made by using a curve which really "fits" the means or by using the means themselves. The latter procedure would entail an error of estimate which would be a function of the variance within the arrays about the array means. An over-all variance about the means of the vertical arrays can be calculated by squaring the deviations about the mean of each array, summing these for all arrays, and then dividing by  $N$ . The resulting variance for the vertical arrays may be labeled  $\sigma_{ay}^2$ , for the horizontal arrays,  $\sigma_{ax}^2$ .

The *correlation ratio*,  $\eta$ , in terms of the accuracy with which  $Y$ 's can be predicted from  $X$ 's is defined as

$$\eta_{yx}^2 = 1 - \frac{\sigma_{ay}^2}{\sigma_y^2} \quad (84)$$

and for  $X$ 's predicted from  $Y$ 's, we have

$$\eta_{xy}^2 = 1 - \frac{\sigma_{ax}^2}{\sigma_x^2} \quad (84a)$$

Are two  $\eta$ 's necessary? We have not proved herein that the variance about the mean is smaller than about any other point, but this fact is readily deducible from the computational formula for  $\sigma$  in terms of deviations from an arbitrary origin. If  $AO$  coincides with the mean,  $\Sigma d^2$  will equal  $\Sigma x^2$ ; if  $AO$  does not coincide with the mean, a subtractive term will always be involved. It follows that  $\sigma_{ay}$  will be less than  $\sigma_{y \cdot x}$  and that  $\sigma_{ax}$  will be less than  $\sigma_{x \cdot y}$ ; hence both  $\eta$ 's will exceed  $r$ , but to varying degrees, depending upon the extent to which the array means fail to fall on a

straight line. Since it is possible, and likely, that the means for the vertical arrays will not exhibit the same departure from linearity as those for the horizontal arrays, it is not reasonable to expect the two  $\eta$ 's to agree.

The  $\eta$ 's indicate the relative accuracy with which one can predict on the basis of array means, and accordingly they are useful measures of the extent of correlation when the regressions are curvilinear. The correlation ratio can also be utilized when the regression is linear; hence it is more generally applicable than the product moment coefficient, which is useful only in the special case where the assumption of linearity is tenable. The correlation ratio, however, does not enter into the regression equation constants.

Even if the regressions were exactly linear for some defined population, a given sample would show deviations from linearity, and therefore  $\eta$ 's for successive samples would show chance sampling deviations from  $r$ . By how much must  $\eta$  exceed  $r$  before one suspects curvilinearity? The only adequate statistical test for answering this question involves the analysis of variance technique and hence is postponed to Chapter 15.

Another definition of  $\eta$  can be had by starting with the proposition that the variance  $\sigma_y^2$  can be broken down into components, a predictable and an unpredictable part, or  $\sigma_y^2 = \sigma_{my}^2 + \sigma_{ay}^2$ , in which  $\sigma_{my}^2$  is the variance of the array means weighted for the number of cases in the several arrays. Then we have  $\eta$  defined as  $\eta^2_{yx} = \sigma_{my}^2 / \sigma_y^2$  and also as  $\eta^2_{xy} = \sigma_{mx}^2 / \sigma_x^2$ . These are analogous to  $r^2 = \sigma_{xy}^2 / \sigma_y^2$  and  $r^2 = \sigma_{xy}^2 / \sigma_x^2$ , and accordingly we may interpret  $\eta^2_{yx}$  as the proportion of  $Y$  variance explained by or associated with variation in  $X$ .

Since the  $\eta$ 's are most readily computed by methods to be developed later (pp. 272-274), no illustration will be given here.

### RANK CORRELATION

When no measuring instrument is available for a trait, resort is frequently made to rank-ordering by judges. One measure of relationship between variables for which we have individuals ranked is given by  $\rho$  (rho), the Spearman rank-difference correlation coefficient:

$$\rho = 1 - \frac{6\sum D^2}{N(N^2 - 1)} \quad (85)$$

in which  $D$  is the difference between an individual's 2 ranks (for the 2 traits). When we have ranks for one variable and scores for the other we can use the scores as a basis for setting up ranks for the latter, and then compute rho.

Whenever rankings on a given variable involve ties (the judges fail to distinguish between 2 or more individuals or the scores used for ranking are such that 2 or more persons have the same score), the ranks are split between individuals who are in tie positions. Suppose 3 ranks have been assigned and that 2 individuals are tied for the fourth position. If they were distinguishable, they would use up ranks 4 and 5, so we assign each a value of 4.5. Had 3 persons tied for this position, we would split ranks 4, 5, and 6, giving each a rank of 5. Then when we proceed to the remaining individuals we must remember that rank position 6 has been used.

The computation of rho is illustrated in Table 23. The fact that

Table 23. COMPUTATION OF RANK-DIFFERENCE CORRELATION COEFFICIENT, RHO

Persons	Ranks		Differences	
	1st	2nd	$D$	$D^2$
A	3	1	2	4
B	4	2	2	4
C	10	10	0	0
D	8	4.5	3.5	12.25
E	5	6	-1	1
F	9	11	-2	4
G	1	3	-2	4
H	2	7	-5	25
I	13	13	0	0
J	11	4.5	6.5	42.25
K	7	8.5	-1.5	2.25
L	6	8.5	-2.5	6.25
M	12	12	0	0
			0	105.00 = $\Sigma D^2$

$$\rho = 1 - \frac{6(105)}{13(169 - 1)} = .71$$

the algebraic sum of the  $D$ 's must be zero can be utilized as a means of checking the  $D$ -column values.

Rho for ranks based on scores for 2 normally distributed variables tends to be slightly (less than .02) lower than the product



moment  $r$  computed from the scores; hence  $\rho$  is comparable with  $r$  as a measure of the strength of relationship.

To test the significance of  $\rho$ , for  $N$  of 10 or more, we may safely use

$$t = \rho \sqrt{\frac{N - 2}{1 - \rho^2}}$$

which approximates the  $t$  distribution with  $N - 2$  degrees of freedom.

$\rho$  does not possess the mathematical advantages inherent in  $r$ , and therefore has merit only when the observations on one or both variables are ranks instead of measures. Because of judgmental difficulties in assigning ranks for  $N$  large, rank-order data are apt to be confined to small samples, but for  $N$  less than 10 the  $t$  test of the significance of  $\rho$  is not satisfactory. Kendall § has proposed another measure, designated  $\tau$  (tau), for use with ranks which is superior to  $\rho$  in so far as testing significance is concerned when  $N$  is very small. As a measure of the degree of relationship, tau, like  $\rho$ , has the property of being unity for a perfect relationship; for zero and near zero correlation these 2 measures tend to be alike numerically, but for other degrees of association tau tends to be lower than  $\rho$ —at times only two-thirds the magnitude of  $\rho$ . Thus tau is not comparable with  $\rho$  (and  $r$ ), and furthermore there seems to be no specifiable way of estimating one from the other. For a much more adequate discussion of both tau and  $\rho$ , the reader is referred to Kendall.

### THE DISCRIMINANT FUNCTION

Suppose we have 2 or more variables (measured in a graduated fashion) which we wish to combine into a total score for the purpose of discriminating between 2 groups. The question arises as to how best weight the variables so as to obtain maximum difference between the total score means for the 2 groups. This difference must be considered *relative* to the within-groups variability; otherwise we could easily produce a large numerical difference by the simple operation of summing the scores and multiplying by a large constant, whereas the real purpose is to have score distributions with

§ Kendall, M. G., *Rank correlation methods*, London: Griffin, 1948.

the least amount of overlap for the 2 groups. We want the difference to be maximal relative to the spread of scores within the groups.

The simplest way to determine the weights for the several variables is to compute the  $\beta$ 's, thence the  $B$ 's, as in the multiple regression problem. For this purpose, the product moment correlations among the 2 or more independent variables are calculated, and the *point* biserial  $r$  is calculated between each independent variable and  $X_1$ , the dependent variable (membership in one or the other of the 2 groups, with one of the groups consistently designated as corresponding to the first category for the biserial setup).

Actually, since the problem here is that of ascertaining optimum relative weights rather than fitting a regression plane, the  $A$  of the regression equation need not be calculated nor need we worry about  $\sigma_1$  ( $= \sqrt{p_1 p_2}$  of the biserial setup). The weights may be taken simply as  $\beta_2/\sigma_2$ ,  $\beta_3/\sigma_3$ , etc., all multiplied by a constant so chosen as to have weights which exceed, say, 10—thereby avoiding decimals. Some of the weights may be negative, according to the sign of the corresponding  $\beta$ . If all or a majority of the weights are negative, the signs of all may be reversed. The relationship of the total of the optimally weighted scores to group membership is describable by the multiple  $r$  computed by equation (69a). Such a multiple  $r$  is the point biserial between the total weighted scores and belonging to one or the other of the 2 groups. Or one may compute the weighted scores for all  $N$  cases and then make distributions for the 2 groups separately in order to scrutinize the amount of differentiation (or overlap) present.

## CHAPTER 13

### Frequency Comparison: Chi Square

The quantity chi square ( $\chi^2$ ), defined in the last chapter as

$$\chi^2 = \sum \frac{(O - E)^2}{E} \quad (83)$$

or as the sum of the squared discrepancies, between observed and expected frequencies, each divided by the expected frequency, is a statistic which is very useful in a variety of problems involving frequencies. Let us begin by an examination of what might be expected to happen if a penny were tossed 100 times. The expected frequency for heads is 50, and for tails is also 50. If for a particular series of tosses we secured 55 heads and 45 tails, the discrepancies would be +5 and -5. When these discrepancies are squared, each becomes +25, and dividing each squared discrepancy by the expected value we would have  $.5 + .5 = 1.0$  as the value for  $\chi^2$ . Had we obtained 40 heads and 60 tails, the discrepancies of -10 and +10, when squared and divided by  $E$ , would give  $2 + 2 = 4$  as  $\chi^2$ .

Three things are readily apparent from the above: first, the greater the discrepancy relative to  $E$ , the greater the contribution to  $\chi^2$ ; second, the two parts being summed to obtain  $\chi^2$  are *not independent*—when the absolute discrepancy for heads is known, that for tails can be inferred to be the same; and third, the squaring process means that  $\chi^2$  is always a positive quantity regardless of the direction of the discrepancies. A fourth fact becomes apparent if one recalls what happens when a series of tosses is repeated. The number of heads (or tails) secured will vary from one series of 100 tosses to the next; hence the amount of discrepancy will vary, and therefore the magnitude of  $\chi^2$  will vary from series to series. In other words, successive sampling will yield varying

values for  $\chi^2$ . If we knew the sampling distribution for  $\chi^2$ , we could specify the probability of securing by chance as large a value as any obtained  $\chi^2$ , and thereby we could judge whether a given amount of discrepancy is significantly large enough to warrant the conclusion that the coin is biased.

Situations similar to this arise in research work. We may, on the basis of a hypothesis that a certain proportion of individuals possess a given characteristic, state how many of a sample of  $N$  cases would be expected to show the characteristic. Observations on  $N$  cases will provide an observed number. If the hypothesis is tenable, the discrepancy between observed and expected should be no larger than might arise on the basis of chance. If the obtained discrepancy is too large, i.e., not apt to arise by chance, the hypothesis becomes suspect. The student who recalls that the standard error of a proportion can be used in comparing observed with expected proportions may wonder whether another technique is necessary. The answer will be forthcoming.

### CHI SQUARE AND THE BINOMIAL DISTRIBUTION

Perhaps some insight regarding the sampling distribution of  $\chi^2$  can be obtained by a re-examination of the binomial distribution, which was discussed in Chapter 5. Suppose we consider the binomial distribution,  $(p + q)^{10}$  with  $p = q = 1/2$ , as yielding the chance distribution of number of heads when 10 unbiased coins are tossed (see Table 24). When 10 coins are tossed we expect to get

Table 24. THE BINOMIAL AND  $\chi^2$  WHEN 10 COINS ARE TOSSED

Number of Heads	$f$	$\chi^2$	$f$ for $\chi^2$
10	1	10.0	2
9	10	6.4	20
8	45	3.6	90
7	120	1.6	240
6	210	0.4	420
5	252	0.0	252
4	210	0.4	
3	120	1.6	
2	45	3.6	
1	10	6.4	
0	1	10.0	
	<hr/>		<hr/>
	1024		1024

5 heads and 5 tails, that is, the  $E$ 's are 5 and 5, but for a particular toss we will have an observed number of heads (and tails) which may differ from 5 and 5. The observed values, or  $O$ 's, could be 10 heads and zero tails; 9 heads, 1 tail; and so on to zero heads, 10 tails. If we obtained 9 heads and 1 tail, we could write  $\chi^2 = (9 - 5)^2/5 + (1 - 5)^2/5 = 6.4$ . Similarly, if we compute  $\chi^2$  for 10 heads and no tails we get a value of 10.0; for 8 heads and 2 tails we get 3.6; etc. Note that for each  $\chi^2$ ,  $\Sigma E = \Sigma O = 10$ .

The third column of Table 24 gives the values of  $\chi^2$  for various possible sets of observed frequencies for number of heads and tails. All the given numerical values of  $\chi^2$ , except 0, appear twice: 9 tails and 1 head will obviously lead to the same  $\chi^2$  as 9 heads and 1 tail. Now the probability of obtaining 9 heads and 1 tail is  $10/1024$  and the  $P$  for 1 head and 9 tails is also  $10/1024$ ; hence the  $P$  for obtaining a  $\chi^2$  of 6.4 is  $20/1024$ . Likewise, we may combine the appropriate binomially derived chance frequencies ( $f_b$ ) so as to write the chance frequencies for the several  $\chi^2$  values. These appear as the fourth column of the table. We have thus established the chance or probability distribution of  $\chi^2$  for a specified coin tossing situation. A plot of these frequencies against the  $\chi^2$  values will reveal a highly skewed distribution.

The probability of a  $\chi^2$  as large as 6.4 will be  $20/1024 + 2/1024$ , or  $22/1024$ , a value which obviously represents the probability of a discrepancy, between  $O$  and  $E$ , as great as 4 in either direction (at least 9 heads or at least 9 tails). The  $P$  of  $22/1024$  involves 1 tail of the distribution of  $\chi^2$  values, but both tails of the binomial contribute thereto. This fact will need to be recalled below when we discuss one- vs. two-tailed tests of hypotheses.

Before we leave Table 24, it might be well to point out a connection between  $\chi^2$  and  $x/\sigma$ . Consider again an obtained frequency of 9 heads. If we express 9 as a deviation from the mean of the binomial,  $np = 5$ , relative to the  $\sigma$  of the binomial,  $\sqrt{npq} = 1.581$ , we have  $4/1.581$ , which when squared gives 6.401 or the corresponding value of  $\chi^2$  (within limits of rounding error). This agreement is not accidental; as will be seen shortly, under specifiable conditions  $\chi^2 = (x/\sigma)^2 = (CR)^2$ . Another characteristic of  $\chi^2$  is obvious from Table 24: for the 10 coin situation no values of  $\chi^2$  other than those given can be obtained because the possible number of heads (and tails) is a discrete series. This lack of continuity imposes a restriction on the use of  $\chi^2$  which will receive more attention as we proceed.

The  $\chi^2$  values in Table 24 are for possible discrepancies of observed frequencies from an expected frequency of 5 for a *single* toss of 10 coins. Suppose that we have, as shown in Table 25, an

Table 25.  $\chi^2$  FOR DISCREPANCIES OF EXPECTED AND OBSERVED FREQUENCIES WHEN 7 COINS WERE TOSSED 1000 TIMES

Number of Heads	$E$	$O$	$O - E$	$\frac{(O - E)^2}{E}$
7	8	4	-4	2.00
6	55	55	0	.00
5	164	157	-7	.30
4	273	283	10	.37
3	273	267	-6	.13
2	164	177	13	1.03
1	55	45	-10	1.82
0	8	12	4	2.00
Sums	1000	1000	0	7.65
	( $N$ )	( $N$ )		( $\chi^2$ )

observed distribution of frequencies obtained by tossing 7 coins 1000 times, and that we wish to compare these observed frequencies with those expected on the basis of the binomial expansion. We are not concerned this time with a single toss for which the expectation would be 3.5, but rather with the results expected when a large number of tosses are made. Note that both the  $E$  column and the  $O$  column sum to 1000 (or  $N$ ) and that the  $(O - E)$ 's sum to zero. The several contributions to  $\chi^2$  are given in the last column, which sums to 7.65, or the  $\chi^2$  for the entire table. Two other series of 1000 tosses made by students in the author's classes yielded  $\chi^2$  values of 12.52 and 15.02. Two of these values for  $\chi^2$  are larger than any of the values in Table 24, and one reason for this is the fact that more  $(O - E)^2/E$  terms are being summed—8 such values instead of 2. Thus, the possible magnitude of a  $\chi^2$  would seem to be a function of 2 things: the size of the squared discrepancies (relative to their respective  $E$ 's) and the number of categories or possibilities for discrepancy. Actually, the chance or sampling distribution of  $\chi^2$  is only indirectly a function of the number of discrepancies; it is a direct function of the number of *independent* discrepancies or the *degrees of freedom*, which we shall next discuss.



## DEGREES OF FREEDOM

We have seen that the  $\chi^2$  of 6.4 in Table 24 involves two  $(O - E)^2/E$  values:  $(9 - 5)^2/5$  and  $(1 - 5)^2/5$ , or 2 discrepancies of exactly the same absolute magnitude. This means that the 2 discrepancies are not independent—as soon as one is calculated, the other can be written down at once without any further calculation; hence 1 degree of freedom exists. If we study the data of Table 25, we see that, since the discrepancies must sum to zero, all 8 cannot be independent or vary freely. As soon as 7 are known, the eighth is determined. This means that there are 7 degrees of freedom for this situation. If we were to roll a die 600 times and then compare the observed frequency for 6 spots, 5 spots, etc., with the number expected on the basis of a perfectly homogeneous (unloaded) cube, we would have 5 possible independent discrepancies, or 5 degrees of freedom. In each of these situations the expected frequencies are determinable on the basis of some a priori principle, and the only restriction is that the total expected frequency must be the same as the total observed frequency, i.e.,  $N_E$  must equal  $N_O$ . In all such cases the number of degrees of freedom ( $df$ ) is 1 less than the number of categories.

The  $df$  for other situations in which the  $\chi^2$  technique is applicable will follow the same principles as to the number of independent discrepancies, but not the rule just laid down. Suppose we consider a 2 by 2 or fourfold table such as that given in Table 26

Table 26.  $\chi^2$  AND FOURFOLD TABLE

(Expected frequencies in parentheses)

	No	Yes	Totals
Group 1	50 (40)	50 (60)	100 = $N_1$
Group 2	70 (80)	130 (120)	200 = $N_2$
Totals	120 $N_n$	180 $N_y$	300 = $N$

(which contains fictitious data for purpose of ease in exposition). The expected frequencies are set up on the assumption that there is no difference between the 2 groups (the null hypothesis). If this were the case, we would expect that the 180 yeses would be distributed in the 1 to 2 ratio of the right-hand totals; likewise

the 120 noes. Note that the expected frequencies reading across, i.e., 40 and 60, and 80 and 120, are proportional to the marginal totals at the bottom. In determining the  $df$ , we can observe either of 2 things: first, that all 4 discrepancies have the same absolute value, so that when 1 is known the other 3 can be written down at once; or second, that in setting up the expected frequencies, we are restricted by the requirement that the 2 top-row values must sum to  $N_1$ , the next 2 must sum across to  $N_2$ , the left-hand column must sum to  $N_n$ , and the next column to  $N_y$ ; as soon as the value 40 has been ascertained, the remaining 3 expected values become fixed. Either way we look at the situation, we see that there is but 1 degree of freedom even though there are 4 cells or 4 discrepancies.

The fundamental question is: How many of the discrepancies are independent? In practice this can be answered by determining how many categories or cells can be filled in at will before the others become fixed because of the restrictions imposed. If we turn back to Table 22 of the last chapter (p. 204), we see that the restrictions for a 3 by 3 table are similar to those for a 2 by 2 table: the expected frequencies must add across and down to the observed marginal totals. The student should ponder Table 22 long enough to see that the proper  $df$  is 4. The general rule-of-thumb for ascertaining the degrees of freedom for all contingency-type tables of  $k$  rows and  $l$  columns, where the marginal totals are utilized in setting up the expected frequencies, is to take  $df = (k - 1)(l - 1)$ . Thus for the fourfold table we have  $(2 - 1)(2 - 1) = 1$ , and for the 3 by 3 table,  $(3 - 1)(3 - 1) = 4$ , etc. Such tables need not be square; in fact, very often the psychologist wishes to compare 2 groups on the basis of  $k$  possible responses to a question. For this  $k$  by 2 table, the  $df$  becomes  $(k - 1)(2 - 1)$ , or simply  $k - 1$ .

### SAMPLING DISTRIBUTION OF $\chi^2$

Before discussing further the applications of  $\chi^2$ , we turn again to the sampling distribution of this statistic. It is easy enough to see from the coin tossing situations which we have considered above that chance leads to discrepancies between observed and expected frequencies. In those situations wherein we wish to compare groups, we know from the discussion of sampling in

Chapter 5 that differences in responses or characteristics can and will arise as a result of chance sampling even though the 2 universes do not differ. Likewise, contingency tables involving the possible relationship between 2 categorized variables will yield varying chance values of  $\chi^2$  even though no real association exists. Knowing the chance sampling distribution of  $\chi^2$  for various degrees of freedom, we can specify the probability of obtaining a  $\chi^2$  as large as any value and conclude therefrom, according to the situation, that observations do not agree with hypothesized frequencies or that 2 or more groups differ significantly or that a real association exists.

We have already suggested that, for 1 degree of freedom, the distribution of  $\chi^2$  is the same as for  $(x/\sigma)^2$ . The general equation for the  $\chi^2$  distribution \* involves an  $n$  or the  $df$ , and therefore there is no one  $\chi^2$  distribution but a very large number of distributions, one for each value of  $n$ . It happens that practical work seldom involves more than 30 degrees of freedom, so that we need not concern ourselves with all possible distributions. Curves for the distribution of  $\chi^2$  can be drawn for various  $n$ 's with  $\chi^2$  along the abscissa and the ordinates as the  $y$  values obtained by the equation in the footnote. The area under each curve will be 1 unit, as in the unit normal curve. Figure 14 contains curves for 7 different values of  $n$  or  $df$ , so drawn as to be comparable. Note that the shapes of these curves and their general locations along the abscissa vary with  $n$ .

For  $n = 1$ , or for 1 degree of freedom, the curve starts very high (strictly speaking, it is asymptotic to the ordinate and hence starts at infinity) and drops quite rapidly. For this curve the height or  $y$  value at  $\chi^2 = .16$  is .92 (not shown). At  $\chi^2 = .01$ , the height is more than 4 times greater than .92. By the time we reach a  $\chi^2$  of 1.00, the height is .242 (what  $x/\sigma$  value does this height correspond to when the unit normal curve is considered?). Then the curve trails off until, at  $\chi^2 = 6.25$ , the height is about

$$* \quad y = \frac{1}{2^{n/2} \cdot \Gamma\left(\frac{n}{2}\right)} (\chi^2)^{\frac{n}{2}-2} e^{-\frac{\chi^2}{2}}$$

in which  $\Gamma$  indicates the gamma function as defined in texts in advanced calculus.

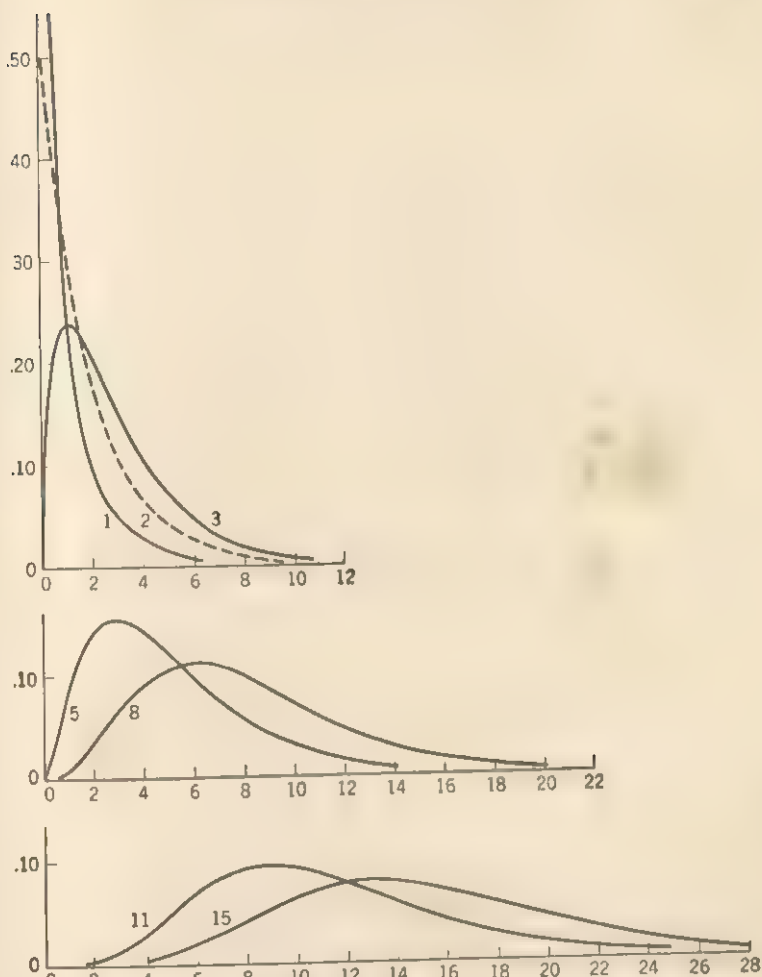


Fig. 14. Chi square distributions for various  $df$ 's  $\chi^2$  along abscissa

.007. Regardless of  $n$ , the right-hand parts of the curves never reach the base line; i.e., they are asymptotic. If we think of the total area under any curve as unity, then the area between ordinates erected at any 2 base-line points, or the area beyond any point, can be expressed as a proportion of the total. Thus, for  $n = 1$ , .99 of the area is beyond (to the right of) a  $\chi^2$  value of .000157, and only .05 is beyond 3.841. Stated differently, the

probability of obtaining a  $\chi^2$  value as large as 3.841 is .05; for  $\chi^2$  as large as 6.635,  $P = .01$ ; and the  $P = .001$  point is at a  $\chi^2$  of 10.827. These hold only for  $df = 1$ .

The curve for  $n = 2$  starts at a height of .50 and then descends, but less rapidly than that for  $n = 1$ . It is readily seen that large values for  $\chi^2$  occur more frequently when  $n = 2$  than when  $n = 1$ . The  $P = .05$  point is at 5.991; i.e., the probability of obtaining by chance a  $\chi^2$  value as great as 5.991 is .05. The .01 point is at 9.210, and the .001 point is at 13.815.

For  $n = 3$ , the distribution curve begins at zero height, rises sharply to a maximum (modal value) at  $\chi^2 = 1$ , and then falls off so that the  $P = .01$  point is at  $\chi^2 = 11.341$ . As  $n$  is taken larger and larger, the distributions become less and less skewed and move farther and farther to the right. The mean of a given distribution always corresponds to a  $\chi^2$  equal to  $n$ , and except for  $n = 1$  the modal value is at a  $\chi^2$  of  $n - 2$ .

The distributions of  $\chi^2$  for varying  $n$ 's are theoretical probability distributions. They may be interpreted as random sampling distributions, and by them one can judge the statistical significance of discrepancies. Their use is exactly analogous to testing the significance of the difference between means, which it will be recalled involves setting up the null hypothesis: if there is no real difference between 2 universe means, the  $D/\sigma_D$  values for successive samples will form a normal curve with center at zero and with unit variance. If a found difference is 1.96 times its standard error, the null hypothesis becomes suspect; if 2.58 times its standard error, the hypothesis of no difference can fairly safely be rejected; if  $D/\sigma_D = 3.00$ , rejection is more definitely indicated. These 3 CR's, it will be recalled, correspond to the .05, the .01, and the .003 levels of significance, for two-tailed tests.

Now  $\chi^2$  can likewise be used to test the null hypothesis. The essential difference between the  $D/\sigma_D$  and the  $\chi^2$  techniques is that the latter involves skewed probability distributions; but, knowing the distribution for a given  $n$ , one can ascertain the necessary value of  $\chi^2$  for the .05, the .01, the .001, or other levels of significance. The statement of the null hypothesis in connection with  $\chi^2$  may vary slightly according to the given situation. If the frequencies in the universe agree with the a priori expected frequencies, if the frequencies in 2 or more universes are the same, if there is zero association in the universe between 2 classi-

fications or variables—if any such conditions hold for the universe or universes, then successive samplings will yield  $\chi^2$  values which will distribute themselves in a determinable manner, thus permitting one to specify the probability of obtaining by chance a  $\chi^2$  value as large as any given or obtained value. When this probability is small, say .01 or less, the null hypothesis is rejected, and its rejection implies that there are real discrepancies or real differences exist or there is a real association.

Since the random sampling distribution of  $\chi^2$  depends upon the  $df$ , which varies from situation to situation, it is not feasible to give a rule-of-thumb criterion in terms of the magnitude of  $\chi^2$  which would be deemed significant. If we adopt  $P = .01$  as the level of significance we wish to attain, then we need to refer to available tables of  $\chi^2$  in order to find how large  $\chi^2$  must be to correspond to this level; likewise for any other chosen level of significance. Probability tables for  $\chi^2$  are available in 2 forms. One form, Fisher's (see Table D of the Appendix), gives the values of  $\chi^2$  which will be exceeded by chance a specified number of times, such as .10, .05, .01, and .001. Elderton's table,<sup>†</sup> gives the probabilities for obtaining chi squares as large as specified values expressed as integers, such as 1, 2, 3 ..., 21, 22. Both tables include varying degrees of freedom. Because of an early erroneous notion as to the meaning of degrees of freedom, Elderton's table must be entered with  $df$  equal to 1 less than his  $n'$  values, e.g., use  $n' = 4$  when  $n$  or  $df = 3$ . Elderton's table has one advantage over that given in our Appendix:  $P$  values as small as .000001 can be ascertained.

For  $n$ 's larger than 30, the expression  $\sqrt{2\chi^2} - \sqrt{2n - 1}$  will have a sampling distribution which will follow very closely the unit normal curve. The probability is accordingly .05 that this expression will exceed +1.64, and .01 that it will exceed +2.33, by chance.

Before the possible applications of  $\chi^2$  are summarized, a word should be said about the underlying assumptions which restrict its usage. The probability figures in the tables of  $\chi^2$  are based on continuous distributions, whereas, as pointed out earlier, the chi squares calculated in practice form a discrete series. It is assumed

<sup>†</sup> Table XII in Pearson, Karl, *Tables for statisticians and biometricians, part I*, Cambridge: Cambridge University Press, 1931.



that the distribution of the latter can be approximated by the former. This is similar to approximating the binomial by the normal curve. A second assumption is that the sampling distribution of the observed frequencies about a given  $E$  follows the normal curve. One can seldom, if ever, check on the tenability of this assumption, but it is possible to specify conditions where the assumption will not hold. If any one  $E$  is small, it is not possible to have a normal distribution of  $O$ 's about it even though the total  $N$  is large. For instance, if  $E = 2$ , the  $O$ 's are restricted on one side of  $E$  to zero and 1, whereas on the other side the possible values run 3, 4, 5, and upward. Such a curtailment ordinarily leads to a skewed distribution for the observed frequencies. Now it is obvious that, when  $E$  is small, we have a greater amount of discontinuity; hence the sampling distribution of observed frequencies will be discrete instead of continuous as called for by the normal curve. It would seem, therefore, that small expected frequencies lead to a violation of both the fundamental assumptions underlying the use of  $\chi^2$ . Various criteria have been proposed for the required size of  $E$ . Some say that the  $\chi^2$  technique is inapplicable when any one  $E$  is less than 10; others say that an  $E$  may be as small as 5. We would suggest that, when possible, adjacent categories be combined so as to have no  $E$  less than 10; if such a combination is impossible and an  $E$  is less than 10 but greater than 5,  $\chi^2$  may be used, providing one is cautious as to the conclusions drawn therefrom. A correction for continuity when  $df$  is 1, as in a fourfold table, is available and will be given later.

A third assumption is that the observations be independent of one another. This assumption is violated when the total of the observed frequencies exceeds the total number of persons in the sample(s). Such an inflation of  $N$  occurs when multiple observations are made on each person and each person is counted more than once (cf. p. 99).

### APPLICATIONS

The chief situations for which it is permissible to use  $\chi^2$  may be classified into 3 types.

1. The discrepancy of observed frequencies from frequencies expected on the basis of some a priori principle. Such situations

are most frequently found in genetics, wherein it is hypothesized that certain crossings should lead to the presence, in a certain proportion of offspring, of some defined characteristic or variation thereof. The frequency table for such situations is 1 by  $k$ , with  $k - 1$  degrees of freedom, since the only restriction is that the expected frequencies must sum to  $N$ . This type of situation does not arise often in research in the social sciences.

2. Contingency tables. Here we have 2 types of situations which differ only in the methods of classifying

a. We may have a contingency table which is analogous to a correlation table in that both classifications are based on continuous or ordered discrete variables for which we have only categorized information for  $N$  individuals. The 2 variables might be in dichotomy (fourfold table), or one might be a dichotomy and the other manifold, or both might involve multiple categories. For these contingency tables it is meaningful to speak of the correlation between the 2 variables, and the degree of correlation might be appropriately specified by the tetrachoric  $r$  or the fourfold point  $r$  or the contingency coefficient (corrected or uncorrected); which measure is used depends upon meeting the requisite assumptions. In so far as we are concerned only with  $\chi^2$ , we have the means for testing the significance of the correlation or association as a chance departure from zero or no relationship, and the significance test can be used without knowledge of the degree of correlation. Such a test of significance is sometimes spoken of as a test of independence—are the 2 classifications independent? If so,  $\chi^2$  should be no larger than would arise by chance. If we have evidence for correlation or a lack of independence from the  $\chi^2$  technique, we can proceed to calculate an appropriate coefficient for measuring the degree of correlation or the strength of association. The student should, as an exercise, convince himself that  $\chi^2$  *per se* is not a measure of association.

b. The other contingency-type situation involves classification into categories for one variable vs. classification into unordered groups for the other, or one unordered grouping vs. another. The fundamental problem is apt to be that of comparing 2 or more groups with regard to multiple responses; i.e., we want a test of the difference between groups rather than a measure of correlation, which would not be entirely meaningful except in the loose sense that a particular response is associated more often with a

particular group. As previously stated, the  $df$  for a  $k$  by  $l$  contingency table is  $(k - 1)(l - 1)$ .

3. Goodness of fit. If we wish to check on whether it is reasonable to believe that a given frequency distribution is, within the limits of chance sampling, of the normal or some other specified type, a frequency curve having the same basic constants (e.g.,  $N$ ,  $M$ , and  $\sigma$  for the normal curve) as those computed from the observed frequency distribution can be fitted to the data. If a normal curve is being fitted, the table of normal curve functions is used to set up the theoretical or expected frequencies for the several grouping intervals. Then  $\chi^2$  can be computed in the usual manner. The  $df$  will correspond to the number ( $k$ ) of grouping intervals less the number of constants derived from the data and used in the fitting process. For the normal curve the observed and theoretical distributions are made to agree as to  $N$ ,  $M$ , and  $\sigma$ ; hence  $df = k - 3$ . An attempt will be made later to explain the reasoning back of the determination of  $df$  when checking the goodness of fit of frequency curves.

**Fourfold contingency tables.** For illustrative purposes, let us first apply  $\chi^2$  to a couple of 2 by 2 contingency tables for which the tetrachoric  $r$ , as well as the contingency coefficient, is an appropriate measure of the degree of correlation. Before we do this, it might be well to recall that  $\chi^2$  for a fourfold table can be computed by a simple formula which does not require calculation of the 4 expected frequencies. Let the fourfold frequencies and marginal totals be set up as in Table 27. Chi square can be com-

Table 27. SETUP FOR COMPUTING  $\chi^2$  FROM A FOURFOLD TABLE BY MEANS OF A FORMULA

A	B	A + B
C	D	C + D
A + C	B + D	N

puted from

$$\chi^2 = \frac{N(AD - BC)^2}{(A + B)(C + D)(A + C)(B + D)} \quad (86)$$

This is simpler than calculation from the discrepancies between observed and expected frequencies. The requisite that no *expected*

frequency shall be less than 5 still holds. A quick check on this can be obtained by multiplying the smaller right-hand marginal frequency by the smaller frequency on the bottom margin and dividing the product by  $N$ . This will yield the smallest expected frequency. In Table 28 will be found 2 fourfold tables for Stanford-

Table 28.  $\chi^2$  APPLIED TO CONTINGENCY (FOURFOLD) TABLES

		Item 1				Item 3			
		-	+			-	+		
Item 2	+	29	39	68	Item 4	+	34	37	71
	-	22	10	32		-	94	35	129
		51	49	100			128	72	200
		$\chi^2 = 5.93$					$\chi^2 = 12.40$		
		$P$ about .01					$P$ less than .001		

Binet items. Direct substitution into formula (86) yields the 2 chi squares at the bottom of the table. The  $P$  values are approximately .01 and less than .001, respectively. We can be reasonably sure that there is some correlation between the first 2 items, and fairly certain that items 3 and 4 are correlated. The value of the tetrachoric  $r$  is .40 for each table, and the contingency coefficient (with no corrections) is .24 for each table. Thus we see that the  $\chi^2$   $P$ 's associated with the same degree of correlation can be different. Why? Would it be possible for 2 fourfold tables to yield the same  $\chi^2$   $P$ , yet differ in the degree of relationship?

Another application of  $\chi^2$  to fourfold tables is given in Table 29,

Table 29.  $\chi^2$  USED TO TEST SEX DIFFERENCES IN PASSING (+) OR FAILING (-) A BINET ITEM

		6		7		8		9	
		-	+	-	+	-	+	-	+
B		84	18	66	36	58	44	37	66
G		93	8	80	20	62	39	52	49
		177	26	146	56	120	83	89	115
		203		202		203		204	
		$\chi^2$ 4.30		5.89		.43		5.02	
		$P$ <.05		<.02		<.50		<.05	

in which the sexes at 4 age levels are compared in performance on a Stanford-Binet item. None of the  $\chi^2$  values reaches 6.635, the value corresponding to the .01 level of significance, but 3 of them are large enough to suggest a real sex difference. That a real difference may exist is also suggested by the fact that the boys are consistently superior at all 4 age levels. This brings us to an important property of  $\chi^2$ . The several chi squares for independent (i.e., based on different samples) tables may be summed to a total  $\chi^2$ , with  $df$  equal to the sum of the  $df$ 's for the chi squares being summed. Thus for Table 29 we have  $4.30 + 5.89 + .43 + 5.02 = 15.64$  as a  $\chi^2$  based on 4 degrees of freedom, by which we can judge the significance of the over-all sex differences shown in the 4 tables. With  $\chi^2 = 15.64$  and  $n = 4$ , we find (from Table D) that  $P$  is less than .01 (for  $n = 4$ , a  $\chi^2$  of 13.28 corresponds to the .01 level). If one turns to Elderton's tables, it can be ascertained that  $P$  is about .004. In other words, as great a sex difference, considering all 4 age groups, would arise 4 times in 1000 by chance; hence it would be concluded that a real difference does exist for this item.

This combinatorial property of  $\chi^2$  is important for all situations where frequency data from different groups cannot first be legitimately combined because of age or other differences. It is most useful when consistency is present among several comparisons, none of which taken singly possesses statistical significance. However, neither consistency nor insignificance for single comparisons constitutes a requisite for using the sum of chi squares as an over-all test of significance or as a means of arriving at 1 summary probability figure.

The single age comparisons in the above example could, of course, be made by means of proportions. This could be done by formula (21) of Chapter 5, the discussion of which (pp. 60-61) should be reviewed at this time. Let us examine the connection between the  $\chi^2$  technique and the  $D \sigma_D$  for proportions method of testing the significance of the difference between 2 groups, the individuals of which have been classified as either passing or failing, saying either yes or no, possessing or not possessing a characteristic, etc. All such comparisons begin with a fourfold frequency table of the type symbolized in Table 27, or an equivalent (the frequencies may have been recorded for only 1 category of the dichotomy, say the yeses, from which the frequencies for the

other category may be readily inferred by subtraction). Table 30 contains the basic table of frequencies for the presence (+) or absence (-) of a characteristic for groups 1 and 2, and the basic

Table 30. SCHEMA FOR COMPARING GROUPS VIA  $\chi^2$  AND VIA DIFFERENCE BETWEEN PROPORTIONS (OR PERCENTAGES)

		Frequencies		
		+	-	
Group	1	A	B	$A + B = N_1$
	2	C	D	$C + D = N_2$
		$A + C$	$B + D$	$N$
Proportions				
		+	-	
Group	1	$p_1 = A/N_1$	$q_1 = B/N_1$	$p_1 + q_1 = 1.0$
	2	$p_2 = C/N_2$	$q_2 = D/N_2$	$p_2 + q_2 = 1.0$
		$p = (A + C)/N$	$q = (B + D)/N$	$p + q = 1.0$

table of proportions obtained by dividing the frequencies by the proper  $N$ 's is indicated. Note that the  $p$  and  $q$  values on the bottom margin are the proportions to use in formula (21) for the standard error of the difference between  $p_1$  and  $p_2$ . Note also that  $p_1 = A/N_1 = .1$  ( $A + B$ ) and that  $p_2 = C/N_2 = .1$  ( $C + D$ ).

In order to avoid carrying along a square root sign or radical, and for another reason which if not now obvious will soon become so, let us write the square of the expression for the critical ratio of the difference between the two proportions,  $p_1$  and  $p_2$ , thus,

$$\frac{D^2}{\sigma^2_D} = \frac{(p_1 - p_2)^2}{\frac{pq}{N_1} + \frac{pq}{N_2}}$$

When we replace all the proportions by their equivalents involving frequencies and the proper  $N$ 's and also substitute frequencies for  $N_1$  and  $N_2$ , we have



$$\begin{aligned}
 \frac{D^2}{\sigma_D^2} &= \frac{[A/(A+B) - C/(C+D)]^2}{\frac{[(A+C)/N] \cdot [(B+D)/N]}{A+B} + \frac{[(A+C)/N] \cdot [(B+D)/N]}{C+D}} \\
 &= \frac{(AC + AD - AC - BC)^2}{[(A+B)(C+D)]^2} \\
 &= \frac{(A+C)(B+D)(C+D) + (A+C)(B+D)(A+B)}{N^2(A+B)(C+D)} \\
 &= \frac{(AD - BC)^2 N^2}{\{(A+B)(C+D)[(A+C)(B+D)(C+D)] \\
 &\quad + (A+C)(B+D)(A+B)\}} \\
 &= \frac{(AD - BC)^2 N^2}{(A+B)(C+D)(A+C)(B+D)(A+B+C+D)} \\
 \frac{D^2}{\sigma_D^2} &= \frac{(AD - BC)^2 N}{(A+B)(C+D)(A+C)(B+D)}
 \end{aligned}$$

which equals  $\chi^2$  as given by formula (86) for the fourfold table. This confirms a fact already mentioned, that for 1 degree of freedom  $\chi^2$  is the same as the square of the critical ratio. Since formula (21) is applicable only for comparing proportions based on independent samples, it follows that  $\chi^2$  is similarly restricted. That is,  $\chi^2$  as computed from a fourfold table by (86) does not allow for any correlational factor which might be introduced because the 2 groups consist of paired or matched individuals or for the correlational factor which would be present if  $p_1$  and  $p_2$  (or the corresponding frequencies) were based on the *same* individuals as in a pretest, intervening experience, posttest situation.

**Significance of changes.** The student should carefully note that although the application of  $\chi^2$  to fourfold tables of frequencies like that of Table 6 in Chapter 5, which is here reproduced with minor changes as Table 31, provides a means of testing the significance of the association or correlation between 2 sets of responses, such an application does not test the significance of change from the first to the second set of responses. This latter test can

Table 31. FOURFOLD TABLE OF FREQUENCIES AND PROPORTIONS FOR A FIRST SET VS. A SECOND SET OF RESPONSES FROM THE *Same* INDIVIDUALS

		Frequencies				Proportions			
		2nd				2nd			
		-	+			-	+		
1st	+	A	B	A + B	1st	a	b	p <sub>1</sub>	1.0
	-	C	D	C + D		c	d	q <sub>1</sub>	
		A + C	B + D	N			q <sub>2</sub>	p <sub>2</sub>	

be made by means of formula (20) of Chapter 5, p. 59. It is also possible to test the significance of any found change by the use of  $\chi^2$ . To do this, we first note that a net change for the group must necessarily involve the difference between the frequencies,  $A$  and  $D$ , since the  $B$  and  $C$  cases represent those who showed no change. The null hypothesis would be that the universe frequencies are not different; i.e., for a given sample,  $A$  and  $D$  would differ only as a result of chance sampling. Since  $A + D$  represents the total number of individuals who changed (the  $A$ 's from + to -, and the  $D$ 's from - to +), in setting up the null hypothesis concerning the net change it would seem appropriate to say that, if  $A + D$  individuals changed,  $(A + D)/2$  would change in one direction and  $(A + D)/2$  in the other direction. Thus  $(A + D)/2$  would become the expected frequency; then  $A - (A + D)/2$  and  $D - (A + D)/2$  would become the discrepancies between observed and expected (on the basis of the null hypothesis) frequencies. If  $A = D$ , both discrepancies would become zero. Squaring each discrepancy and dividing by  $E$  and then summing the 2 quotients or doubling either one will give a  $\chi^2$  which is based on 1 degree of freedom (why 1 degree of freedom?). A little algebraic manipulation shows that

$$\chi^2 = \frac{(A - D)^2}{A + D} \quad (87)$$

for the particular situation in which we wish to test the significance of over-all changes.

Comparison of formula (87) with formula (19a), p. 58, shows that we again have a  $\chi^2$ , with 1 degree of freedom, which equals

the square of an  $x/\sigma$ , or critical ratio. The reasoning back of the statement given on p. 60 that formulas (19a), (19b), and (20) are inapplicable unless  $A + D$  equals 10 or more should now be clearer to the reader. If  $A + D$  were less than 10, the two  $E$ 's would be less than 5, an acceptable though none too conservative lower limit for  $E$ . A correction (for continuity) needed when the  $E$ 's are smaller than 10 will be given shortly. One thing which may puzzle the reader at this time is the fact that formula (87) does not contain a total  $N$ . Its algebraic equivalent,  $(D'\sigma_D)^2$ , with  $\sigma_D$  calculated by formula (20), does contain  $N$ , so the absence of  $N$  from (87) is more apparent than real.

The advantage of the  $\chi^2$  over the  $D'\sigma_D$  technique for testing the significance of net changes in responses lies in the fact that  $\chi^2$  values for 2 or more groups which have been used in an experiment can be summed to a new  $\chi^2$  with  $n$  equal to the sum of the separate  $df$ 's; in this case  $n$  equals the number of chi squares being summed.

Formula (87) is, of course, not restricted to situations involving changes in responses. If we have the same individuals giving, say, yes or no responses to 2 different questions and we desire to test the significance of the difference between the frequencies (or proportions) of yeses or noes, formula (87) is applicable. Or suppose we wish to know whether there is a significant difference in the difficulty of 2 test items which have been administered to the same group. For example, in Table 28 we have 49 and 68 individuals passing items 1 and 2 respectively. Since  $N = 100$ , the proportions are .49 and .68 (or 49 and 68 per cent). By formula (87) we have  $\chi^2 = (29 - 10)^2 / (29 + 10) = 9.26$ , which for 1 degree of freedom falls between the .01 and .001 levels of significance; hence it would be concluded that the 2 items are different in difficulty. If we use formula (20), we get a critical ratio,  $(p_2 - p_1)/\sigma_D = (.68 - .49)/\sqrt{(.10 + .29)/100} = .19/.0624 = 3.04$ , which leads to the same probability figure as that for a  $\chi^2$  of 9.26. Either method may be used. Both make due allowance for the correlation which is present because the frequencies or proportions being compared are based on the *same* individuals.

**Correction for continuity.** We have already pointed out that, since the sampling distribution of  $\chi^2$  is continuous, the use of  $\chi^2$  when any one  $E$  is less than 5 is questionable. For fourfold contingency tables, an allowance for discontinuity can be made by

applying Yates's correction for continuity, which should always be used when any one  $E$  in such a table is less than 5 and is advisable when an  $E$  is less than 10. A small  $E$  is most likely to occur either when the total  $N$  is small or when one or both of the marginal totals involve extreme dichotomies. It is easy to determine the smallest  $E$  by dividing the product of the 2 smaller marginal frequencies by the total  $N$ . Yates's correction can be incorporated in formula (86), which becomes

$$\chi^2 = \frac{N(|AD - BC| - N/2)^2}{(A + B)(C + D)(A + C)(B + D)} \quad (86a)$$

and indicates that the absolute difference between  $AD$  and  $BC$  is to be reduced by  $N/2$ . Formula (87) can also be written to include a correction for continuity. The corrected form

$$\chi^2 = \frac{(|A - D| - 1)^2}{A + D} \quad (87a)$$

involves decreasing the absolute value of the difference between  $A$  and  $D$  by 1. Formula (87a) is to be preferred to (87) when  $A + D$  is less than 20. The reasoning back of Yates's correction is precisely the same as that given on p. 48 of Chapter 5.

**One-tailed vs. two-tailed test.** It will be recalled from our discussion of the sampling distribution of  $\chi^2$  that the  $P$ 's obtainable from Table D are the probabilities of the chance occurrence of as large a  $\chi^2$  as that observed; that is, levels of significance such as  $P = .05$  or  $.01$  or  $.001$  are based on *one* (the right-hand) tail of the sampling distribution of  $\chi^2$ . Does this mean that it is a one-tailed test in the hypothesis testing sense discussed earlier (pp. 62-64)? Let us recall a couple of facts. First, when using the binomial to indicate something of the nature of the  $\chi^2$  distribution we saw that both tails of the binomial were combined as 1 tail of the  $\chi^2$  distribution. Second, for 1 degree of freedom  $\chi^2 = (x/\sigma)^2$ . Now an  $x/\sigma$  of 1.96 corresponds to the  $P = .05$  level as a two-tailed test. The square of 1.96 gives a  $\chi^2$  of 3.84, which we can see from Table D also corresponds to the .05 level. Hence the  $P$ 's, for 1 degree of freedom, read from Table D are equivalent to those based on the two-tailed test despite the fact that only 1 tail of the  $\chi^2$  distribution is involved.

If the decision to be made or the hypothesis to be tested calls for a one-tailed test, the  $P$ 's from Table D need to be halved: a  $\chi^2$  of 5.41 (instead of 6.64) is required for the .01 level, and a  $\chi^2$  of 2.71 (instead of 3.84) gives the .05 level. Incidentally, for 1 degree of freedom, a  $\chi^2 P$  can, obviously, be obtained by entering its square root into the normal curve table—whether such a  $P$  from  $x/\sigma$  is based on one or both tails of the normal distribution depends on the hypothesis being tested. As we proceed, the student should convince himself that the notion of direction of differences, hence the idea of a one-tailed test, doesn't make sense in other applications of  $\chi^2$ .

**Comparison of two or more correlated proportions.** Formula (87) has recently ‡ been extended to provide a method for testing whether 3 or more nonindependent proportions (or sets of frequencies) differ significantly among themselves. For example, we may have pass-fail (or yes-no, or some other dichotomous) information on  $C$  items (or questions) for  $N$  individuals; or we may have only 1 item with responses from  $N$  persons under  $C$  different conditions; or 1 item with responses from  $N$  sets of  $C$  matched persons each, that is,  $C$  matched groups.

Data from such situations can be arranged in a table consisting of  $N$  rows and  $C$  columns. The total number of passes (yeses) in a given column divided by  $N$  will, of course, be the proportion of passes (or yeses) in that column. Do these  $C$  proportions (or the totals) differ significantly in an over-all sense? The null hypothesis is that all the proportions are the same except for chance. To test the null hypothesis we will need to obtain not only the column totals (number of passes) but also a similar total for each of the  $N$  rows. Let  $T$  stand for the total in any column and  $X$  stand for the total in any row. This  $X$  is a sort of "score" for the person—his number of passes (or yeses) on the  $C$  items. Cochran shows that the sampling distribution of the quantity

$$Q = \frac{(C - 1)[C\sum T^2 - (\sum T)^2]}{C\sum X - \sum X^2} \quad (88)$$

follows the  $\chi^2$  distribution with  $C-1$  degrees of freedom for  $N$  large ( $N > 30$ , presumably).

‡ Cochran, W. G., The comparison of percentages in matched samples, *Biometrika*, 1950, 37, 256-266.

The computation of  $Q$  is so easy that it need not be illustrated. If an obtained  $Q$  exceeds the  $\chi^2$  required for a chosen level of significance, one concludes that the (correlated) proportions do differ in an over-all sense, that is, they are not homogeneous. It can be argued that unless  $Q$  is significant one is not justified in singling out the proportions (or columns) which give large differences for the purpose of testing the significance of the difference since such selection tends to capitalize on chance differences.

**Chi square for 2 by  $k$  tables.** The calculation of  $\chi^2$  from a table with 2 rows and  $k$  columns (or 2 columns and  $k$  rows) can be accomplished by way of expected cell frequencies calculated as previously suggested from the marginal totals or by means of

$$\chi^2 = \frac{N^2}{A_i B_i} \left[ \sum \frac{B_i^2}{A_i + B_i} - \frac{B_i^2}{A_i + B_i} \right] \quad (89)$$

in which the  $A$ 's and  $B$ 's have the meanings indicated in Table 32,

Table 32. THE CALCULATION OF  $\chi^2$  FROM A 2 BY  $k$  TABLE: 2 GROUPS AND  $k$  (= 5) RESPONSES

Col. A	Col. B	Col. C	Col. D	Col. E
Group				
I	II	$A_i + B_i$	$\frac{B_i}{A_i + B_i}$	$\frac{B_i^2}{A_i + B_i}$
1 27(= $A_1$ )	15(= $B_1$ )	42	.3571	5.36
2 26(= $A_2$ )	16(= $B_2$ )	42	.3810	6.10
3 247(= $A_3$ )	110(= $B_3$ )	357	.3081	33.89
4 41(= $A_4$ )	8(= $B_4$ )	49	.1633	1.31
5 39(= $A_5$ )	15(= $B_5$ )	54	.2778	4.17
Totals 380(= $A_i$ ) + 164(= $B_i$ ) = 544(= $N$ )			.3015	50.83
				49.44
				1.39

$$\frac{544^2}{(380)(164)} = 4.75; \quad \chi^2 = (4.75)(1.39) = 6.60$$

$$n = 4, \quad P = .16$$

wherein will be found the frequencies for 2 groups classified according to 5 response categories. The necessary computations required by formula (89) are also included in the table. Note that, as usual, the marginal totals are first found by summing across and down. Column D is obtained by dividing the entries in



column B by the adjacent values in column C, and column E results from multiplying the D column values by the B column figures. These same operations, when applied to the last (or totals) line, lead to the column E entry of 49.44, which is the value of the  $B_t^2/(A_t + B_t)$  term in formula (89). Summing the first 5 figures in column E yields 50.83, or the  $\Sigma$  term of (89), and the difference between 50.83 and 49.44 is 1.39, the value of the bracketed part of the formula. When this is multiplied by  $N^2/A_t B_t$ , we have  $\chi^2$ , which for a  $df$  of 4 yields a  $P$  of about .16. In other words, once in 6 trials differences as large as those in Table 32 would occur by chance; hence we have insufficient evidence for concluding that the universes from which these 2 samples were drawn differ in regard to their responses to the asked question.

If one had to depend upon the  $D \sigma_D$  technique for testing the significance of the group differences in Table 32, 5 critical ratios would result for each category there is a possible difference in proportions or percentages with a standard error for each difference. The 5  $CR$ 's might, and usually would, lead to 5 different  $P$  values with a consequent predicament as to interpretation. Off-hand, it might be argued that, if any  $CR$  or  $P$  so determined reached an acceptable level of significance, one would be justified in concluding that the difference between the groups was real rather than chance. That such an argument may be fallacious is well illustrated by the data of Table 32, which are actual data. When these data first came to the author's attention, the table was in percentage form with a  $CR$  worked out only for the category showing the largest difference. This  $CR$ , based on formula (21), was 2.54, which is near the  $P = .01$  level of significance, and it had accordingly been concluded that a real difference had been found. Now, when we consider the  $\chi^2 P$  of .16 for the over-all comparison, we are not justified in placing much confidence in such a conclusion.

Why the apparent inconsistency between 2 tests of significance? Since most investigators are looking for group differences rather than group similarities, there is the tendency to single out a category for comparison not because of intrinsic a priori interest in that category but because it happens to yield the largest difference. By this a posteriori selection one tends to capitalize on differences which may be large mainly as a result of chance. A

similar situation occurs when we have the means for several groups—the largest of the possible differences may be the largest partly or entirely as a result of chance. As will be seen in the discussion of the analysis of variance in Chapter 15, before any one difference is tested, an over-all test of significance should be applied. If this over-all test yields a significant *P*, then and only then is one justified in proceeding to an examination of single categories. Thus the use of  $\chi^2$  for such situations as are exemplified in Table 32 not only provides an over-all single index of significance but also helps us avoid false conclusions.

**Application to *k* by *l* tables.** Consider the data of Table 33,

Table 33. TABLE OF FREQUENCY OF 3 POSSIBLE RESPONSES FOR 3 GROUPS OF INDIVIDUALS—PERCENTAGES IN THE PARENTHESES ADD DOWNWARD TO 100 \*

Motivation of Conscientious Objectors	Group			Total
	I	II	III	
Not cowards	24(27.0)	56(53.8)	71(69.6)	151
Partly cowards	30(33.7)	23(22.1)	19(18.6)	72
Cowards	35(39.3)	25(24.0)	12(11.8)	72
<i>N</i> 's	89(100.0)	104(99.9)	102(100.0)	295

\* Data from Leo Crespi, *J. Psychol.*, 1945, 19, p. 285.

which contains a contingency-type table involving 3 groups and 3 possible opinion responses. To test the significance of the differences between the groups by use of the *CR* technique would involve comparing the percentages for group I vs. II, I vs. III, and II vs. III, for each of the 3 responses—a total of 9 *CR*'s. Even though there is no short-cut formula for computing  $\chi^2$  for such a table, its calculation is far quicker than the determination of 9 *CR*'s. Straightforward computation gives  $\chi^2 = 36.58$ , which for *df* = 4 is double the value of the  $\chi^2$  needed for the *P* = .001 point. From Elderton's table we find that *P* is about .000001; hence Table 33 as a whole exhibits highly significant differences between the groups.

Perhaps a better understanding of the extent of the differences can be had by considering the percentages given in parentheses in the table. Membership in group III means a greater tendency to the "not cowards" response. Group I tends more to give the

"cowards" response. Now it happens that the 3 groups, I, II, and III, can be (and are) placed in an ordered series for amount of education: grammar school, high school, and college respectively. Thus the association shown in the table is in the direction of less disparagement of conscientious objectors by those in the higher educational level. The strength of association or degree of correlation is represented by a contingency coefficient of .33, which may seem rather low in light of the highly significant  $\chi^2 P$ . This illustrates a point which most readers will already have grasped: high statistical significance and a high degree of association are far from synonymous. Consideration of the data of Table 33 readily indicates the difficulty of predicting responses when the extent of association is represented by a  $C$  of .33.

As in the 2 by  $k$  table, so here it is better to calculate an over-all  $\chi^2$  before examining by the  $CR$  technique any of the possible separate comparisons. Unless the  $\chi^2 P$  is significant, it is unwise to proceed with such comparisons.

**Goodness of fit.** The use of  $\chi^2$  in testing the goodness of fit of a theoretical curve to an observed frequency distribution is illustrated in Table 34. One starts with an actual distribution, usually with more grouping intervals than in our example, and the descriptive statistical measures therefor. In fitting the normal curve to the distribution of Table 34, we need  $N$ ,  $M$ , and  $\sigma$ . To set up for each interval the frequency which would hold for the best-fitting normal curve, we go through the tedious process of determining the proportionate area under the theoretical curve for each interval. Once the proportions are known, each is multiplied by  $N$  to secure the expected frequencies. The proportions are ascertained by calculating the  $x/\sigma$  value of the boundary limits of the intervals. For example, the 110-119 interval may be thought of as running from 109.5 to 119.5, since IQ's are rounded to the nearest integer. Then  $(109.5 - 104.56)/16.99 = .2907$  as the  $x/\sigma$  for the lower limit, and  $(119.5 - 104.56)/16.99 = .8793$  as the  $x/\sigma$  for the upper limit of the 110-119 interval. Of course, .8793 is also the lower limit for the 120-129 interval. Now the difference,  $.8793 - .2907 = .5886$ , is the same as  $10/16.99$  or  $i/\sigma$ , which is the interval width expressed in  $x/\sigma$  units. Adding .5886 once to .2907 gives .879 (it is sufficient to retain 3 decimals); adding it twice gives 1.468; and so on. Then subtracting .5886 once from .2907 gives  $-.298$ ; subtracting twice gives  $-.886$ ; etc.

Table 34. GOODNESS OF FIT OF NORMAL CURVE TO STANFORD-BINET IQ'S,  
FORM M

			Proportionate			
IQ	O	$x/\sigma$	Area	E	O - E	$(O - E)^2/E$
160	3					
150	13		.0041	12	4	1.33
		2.645				
140	55		.0158	47	8	1.36
		2.057				
130	120		.0512	152	-32	6.74
		1.468				
120	330		.1186	352	-22	1.38
		.879				
110	610		.1958	582	28	1.35
		.291				
100	719		.2316	688	31	1.40
		-.298				
90	592		.1950	579	13	.29
		-.886				
80	338		.1177	350	-12	.41
		-1.475				
70	130		.0506	150	-20	2.67
		-2.064				
60	48		.0155	46	2	.09
		-2.652				
50	7		.0040	12	0	.00
40	4					
30	1					
<hr/>			<hr/>	<hr/>	<hr/>	<hr/>
2970 = N			.9999	2970	0	17.02 = $\chi^2$
M = 104.56			df = 11 - 3 = 8;		P = .03	
$\sigma = 16.99$						

When the boundary limits in terms of  $x/\sigma$  have been set up, the proportionate area for a given interval is found by using the table of normal curve areas. The 2 top intervals have been combined, and likewise the 3 bottom intervals, so as to have no expected frequencies less than 10. The proportionate areas, .0041 and .0040, represent the areas beyond given points, and the  $E$ 's at top and bottom are the number of cases expected beyond these same points. Note that the sum of the proportions should be unity within limits of rounding errors, and that the sum of the expected frequencies should be the same as the sum of the observed frequencies. Perhaps it is unnecessary to point out that the expected

frequencies form an exactly (within limits of rounding errors and for the given intervals) normal distribution which will yield the same  $M$  and  $\sigma$  as the observed distribution with which we started.

Straightforward calculation gives a  $\chi^2$  of 17.02. With  $df = 11 - 3$  (number of intervals minus the number of constants used in the fitting),  $P = .03$ ; i.e., only 3 times in 100 would as large a  $\chi^2$  arise by chance, or only 3 times in 100 would we get a worse fit if the universe of IQ's were distributed as a normal curve. This would lead one to question whether IQ's, as measured by Form M of the 1937 Revision of the Stanford-Binet, are distributed in the normal curve fashion. The same data with intervals of size 5 give a  $\chi^2 P$  of .003, and the degree of kurtosis (by moments) is thrice its standard error; therefore one can conclude that the observed distribution is not a chance departure from a normal distribution.

Thus the  $\chi^2$  technique provides us with a test by means of which we can judge that the frequencies of a given distribution do not follow the frequencies of a theoretical curve closely enough to be regarded as chance departures therefrom. Note that a smaller value for  $\chi^2$  for the example of Table 34 would not prove that the universe is normal even though the  $P$  were as large as .90 or .95. This would merely indicate that the given data were consistent with the normal distribution. As a matter of fact, so-called excellent fits leading to  $P$ 's of .99 or more are suspect. When  $P = .01$ , it is said that chance sampling would lead to a worse fit only once in 100 times; when  $P = .99$ , it is said that chance sampling would lead to a better fit only once in 100 times. In other words, if  $P$  is between .05 and .01, the hypothesis that the universe distribution is of the normal type (or whatever type was fitted) is questionable; if  $P$  is .01 or less, this hypothesis is rejected; if  $P$  is between .95 and .99, one may suspect the fit as being too good; if  $P$  is .99 or more, one should definitely look for an error in calculation or for some type of restraint on the operation of chance. Too good a fit is as open to question as too poor a fit. If  $P$  is between .05 and .95, the fit is said to be satisfactory.

When one is testing the goodness of fit of frequency curves, the  $df$  depends upon the number of grouping intervals and upon the number of restrictions imposed or the ways in which the expected distribution is made to agree with the observed distribution. The general principle back of the determination of  $df$  for  $\chi^2$  as a test of



fit may be illustrated for the case of testing the goodness of fit of the normal curve. The expected and observed distributions are made to agree with respect to  $N$ ,  $M$ , and  $\sigma$ . Suppose that we have  $k$  grouping intervals and that we let  $f_i$  stand for the frequency in the  $i$ th interval and  $X_i$  for its score value (midpoint), and that  $x_i$  represents the corresponding deviation score value for this midpoint. Then the following equations will hold:

$$f_1 + f_2 + f_3 + \cdots + f_i + \cdots + f_k = N$$

$$f_1X_1 + f_2X_2 + f_3X_3 + \cdots + f_iX_i + \cdots + f_kX_k = NM$$

$$f_1x_1^2 + f_2x_2^2 + f_3x_3^2 + \cdots + f_ix_i^2 + \cdots + f_kx_k^2 = N\sigma^2$$

Now, if all the  $f$  values were known except  $f_1$ ,  $f_2$ , and  $f_3$ , those parts to the right of the  $f_3$  term in the first of these equations could be added numerically. The resulting sum could be shifted to the right of the equality sign and then combined numerically with  $N$ , giving an equation of the type  $f_1 + f_2 + f_3 = A$ , where  $A$  equals  $N$  minus the sum of all the frequencies save the first 3. Likewise, the parts beyond the  $f_3$  term in each of the other 2 equations could be summed numerically, shifted to the right, and combined numerically with the constant,  $NM$  for the second and  $N\sigma^2$  for the third equation.

This procedure will lead to 3 simultaneous equations with  $f_1$ ,  $f_2$ , and  $f_3$  as the unknowns:

$$f_1 + f_2 + f_3 = A$$

$$f_1X_1 + f_2X_2 + f_3X_3 = B \text{ (say)}$$

$$f_1x_1^2 + f_2x_2^2 + f_3x_3^2 = C \text{ (say)}$$

It is a well-known principle of algebra that 3 equations in 3 unknowns will be satisfied (if solvable) by just 1 set of values for the unknowns. For our particular problem, this means that, as soon as the frequencies for all but 3 (any 3) intervals are known, these 3 remaining frequencies are not "free to vary"; they are fixed because of the requirements that the frequencies or functions thereof must add to  $N$ ,  $NM$ , and  $N\sigma^2$ . We accordingly lose 3 degrees of freedom, and therefore when we are testing the fit of a normal curve to a distribution with  $k$  intervals, the  $df$  is  $k - 3$ .

If we wished to ascertain whether the observed distribution of Table 34 could be thought of as a chance departure from a normal



curve with mean equal to 100, the expected frequencies would be so set up as to yield the observed  $\sigma$  and  $N$ , but with  $M = 100$ . The  $df$  would therefore be  $11 - 2$ , since the distributions are forced to agree only as to 2 constants,  $N$  and  $\sigma$ ; hence 2 degrees of freedom are lost.

Chi square can be used to test the significance of the difference between 2 observed frequency distributions, but this simply becomes a 2 by  $k$  table with expected values computed from the marginal totals as previously indicated. In such a situation, it is incorrect to treat either set of frequencies as those expected, against which the other is compared as a set of observed values. Such a procedure does not allow for the fact that both sets of frequencies are subject to sampling fluctuations. If one set of frequencies is for the universe, and the second set is based on a sample from the universe, then the universe frequencies (or proportions) can be used to set up expected frequencies, against which the sample values may be checked in order to test whether the sample represents the universe within the limits of chance sampling error. The  $df$  becomes  $k - 1$ , since this requires only that  $N_E = N_O$ .

In this chapter we have discussed the essential nature of  $\chi^2$  and have pointed out typical applications. By now the student should appreciate the advantages of  $\chi^2$  over percentage comparisons and have some insight into the use of  $\chi^2$  as a means of testing hypotheses.

### EXACT OR DIRECT PROBABILITIES

The  $\chi^2 P$ 's obtainable from Table D are approximations in that areas under a continuous curve are taken as estimates of values which form a point distribution. Even with Yates's correction for continuity, the approximation is none too good when  $E$  values are less than 5. This raises the question as to the criterion for judging the closeness of such approximations, and the answer is that for situations involving 1 degree of freedom it is possible to specify exact probabilities. How?

First, consider the problem of deciding on the basis of a specified number of successes whether a chap can distinguish between 2 cigarette brands. We learned in Chapter 5 that the exact  $P$  for the probability of as many correct identifications can be obtained by the binomial distribution; hence we need not use the normal curve or the  $\chi^2$  approximation. But such approximations are

not only very convenient computationally for  $N$  (or  $n$ ) large, but also are accurate enough. In checking a  $\chi^2 P$  against an exact  $P$  derived from the binomial one must bear in mind the possibility of confusing one- and two-tailed tests; both methods should be alike in this regard.

Second, consider the  $\chi^2$  test of the significance of change (or difference between 2 correlated frequencies or proportions) given by formula (87). An exact  $P$  can be obtained for this situation by resort to the binomial (see p. 58). Again, in calculating the binomial  $P$ , one must give consideration to whether he had intended a one-tailed or a two-tailed test.

Third, consider the fourfold table for which formula (86) is appropriate in testing either the significance of association or the significance of the difference between 2 groups. For this situation the binomial is not applicable (except when the frequencies are equal on one, or both, of the margins). Exact  $P$ 's can be obtained for such tables by a rather tedious procedure which we shall now describe. It can be shown that the probability for a particular observed set of frequencies,  $A, B, C$ , and  $D$ , for fixed margins is

$$P = \frac{(A+B)!(C+D)!(A+C)!(B+D)!}{N!A!B!C!D!}$$

To have a test comparable to the usual significance test we would also need the  $P$ 's for all sets of frequencies deviating farther than the observed set from the null values of no association. This can be made clearer by an example. In Table 35 will be found

Table 35. SERIES OF FOURFOLD TABLE FREQUENCIES REQUIRED FOR CALCULATING  $P$  DIRECTLY AND EXACTLY

I					II					III				
		-	+				-	+				-	+	
+		3	9	12	+		2	10	12	+		1	11	12
-		6	2	8	-		7	1	8	-		8	0	8
		9	11	20			9	11	20			9	11	20

an observed set (part I) and sets showing higher association (parts II and III). Note that each part is derived from the preceding

part by subtracting 1 from both  $A$  and  $D$  and adding 1 to both  $B$  and  $C$ . This process is continued until  $A$  or  $D$  or both become zero. Note that the marginal frequencies remain the same.

Application of the foregoing formula to each table in turn will yield the probability for each set of frequencies, and the sum of these  $P$ 's will be the probability of as great association (in the given direction) as that indicated by the starting (observed) set of frequencies. We have

$$P_I = \frac{(12!)(8!)(11!)(9!)}{(20!)(3!)(9!)(6!)(2!)} = .0367$$

$$P_{II} = \frac{(12!)(8!)(11!)(9!)}{(20!)(2!)(10!)(7!)(1!)} = .0031$$

$$P_{III} = \frac{(12!)(8!)(11!)(9!)}{(20!)(1!)(11!)(8!)(0!)} = .0001$$

The sum of these separate probabilities gives  $P = .0399$ , or .04 (to 2 decimals) as the probability of obtaining sets as extreme (in 1 direction) as the set observed in part I of Table 35. If the situation calls for a two-tailed test,  $P = .0798$ , or .08, as the probability of as large a difference (or as great an association) irrespective of direction. This  $P$  value of .0798 may be compared with a  $\chi^2$   $P$  of .082 when  $\chi^2$  is computed by formula (86a) and with a  $\chi^2$   $P$  of .055 when formula (86) is used. As expected, correction for continuity improves appreciably the estimate of  $P$ .

The computation of the separate  $P$ 's, laborious even with an ordinary table of logarithms, is greatly facilitated by a table of the logarithms of factorials, such as Table XLIX of Part I of Pearson's *Tables for statisticians and biometricians*.

## Comparison of Variabilities

For samples with  $N$ 's greater than 100, the standard error of a standard deviation,  $\sigma_\sigma = \sigma/\sqrt{2N}$ , can be used to test hypotheses with regard to population standard deviations and also can be used in formulas (27a) and (27b), p. 88, to determine the significance of the difference between standard deviations. We have already pointed out the fact that the sampling distribution of  $\sigma$  (and of the unbiased estimate,  $s$ ) is skewed when  $N$  is small; hence we need methods for testing hypotheses about variabilities which make allowance for the skewness of the sample  $\sigma$ 's or  $s$ 's. The student will recall that  $s^2 = \Sigma x^2 / (N - 1)$ , but he may need to refer back to p. 107 for computational procedures. From now on we shall use the symbol  $\hat{\sigma}^2$  in place of  $\sigma_{pop}^2$ .

It can be shown that  $N\sigma^2 \cdot \hat{\sigma}^2$  [exactly equivalent to  $(N - 1)s^2/\hat{\sigma}^2$ ] will, for successive random samples, be distributed as  $\chi^2$  with  $N - 1$  degrees of freedom. This fact permits exact tests of hypotheses regarding a single variance and also provides a method of setting confidence limits for a population variance, but since there seems to be little if any need for such statistical activity in psychology, we shall not elaborate further here. The enterprising student will be able to set up the procedures.

When testing the difference between 2 standard deviations or 2 variances we must, as always, distinguish between situations involving correlated values and situations in which the measures are independent (or based on independent samples). The methods about to be presented are applicable for both small and large samples and are based on differences between variances rather than differences between standard deviations.

**Differences between correlated variances.** Correlated variabilities arise when we have 2 forms of a psychological test

administered to the same group with a  $\sigma$  or  $s$  for each form, or when we have the  $\sigma$  for a first trial vs. the  $\sigma$  for a later trial for the same sample, or  $\sigma$ 's for the performance of 1 group under different experimental conditions, or  $\sigma$ 's based on 2 groups ( $N$  pairs of individuals) related by blood or related by matching. For such situations the difference between variations can be tested by

$$t = \frac{(s_1^2 - s_2^2)\sqrt{N-2}}{\sqrt{4s_1^2s_2^2(1-r_{12}^2)}} \quad (90)$$

or its exact equivalent with  $s_1^2$  and  $s_2^2$  replaced by  $\sigma_1^2$  and  $\sigma_2^2$ . This  $t$  follows the  $t$  distribution with  $N-2$  degrees of freedom.

**Differences between independent variances.** For the purpose of testing the difference between uncorrelated  $\sigma$ 's or  $s$ 's, Professor R. A. Fisher developed the mathematics of the sampling distribution of a function designated by  $z$  and defined as

$$z = \log_e s_1 - \log_e s_2 \quad (91)$$

If successive samples are drawn from a single universe or from 2 universes having the same variance, the sampling variation of  $z$  will center at zero and depend upon  $n_1$  and  $n_2$ , the two  $df$ 's. Note that the sampling distribution is independent of the universe value of the variance or standard deviation. In other words, we do not require an estimate of a standard error which uses information from the samples, as required for the standard error of the difference between  $\sigma$ 's. Probability tables for the  $z$  function are available by which one can, for given  $df$ 's, i.e.,  $n_1$  and  $n_2$ , find how large  $z$  must be for the .05, the .01, and the .001 levels of significance.

The  $z$ , defined by formula (91), has 1 disadvantage: logarithms must be used. Since (91) can be written in the equivalent form

$$z = \frac{1}{2} \log_e \frac{s_1^2}{s_2^2} \quad (92)$$

it is seen that, instead of the difference between 2 logarithms, we have  $z$  as a function of the ratio of the 2 estimated variances. From the sampling distribution of one-half the log of a ratio, the sampling distribution of the ratio itself can be inferred. For  $n_1 = 5$  and  $n_2 = 16$ , the value of  $z$ , which will be exceeded 1 per cent of

the time by chance (the .01 probability level), is .7450. This is one-half the log of the ratio of the 2 variances, and hence the log of the ratio would be 1.4900; by reference to a table of natural logarithms the antilog of 1.4900 is found to be 4.44. That is, as large a ratio as 4.44 would occur .01 time by chance. In order to avoid the necessity of using logs, Professor George W. Snedecor has developed tables for the *variance ratio*, which is defined as

$$F = \frac{s_1^2}{s_2^2} \quad (93)$$

The equation \* of the sampling distribution of  $F$  contains 2  $n$ 's:  $n_1$  for the  $df$  upon which  $s_1$  is based, and  $n_2$  as the  $df$  for  $s_2$ . This means that there is a sampling distribution curve of  $F$  for each possible combination of  $n_1$  and  $n_2$ . The probability table for  $F$  must accordingly be entered with  $n_1$  and  $n_2$  in order to learn what level of significance a given  $F$  reaches. To use Table F of the Appendix, we take the larger of the 2 variance estimates as the numerator in computing  $F$ , and the  $df$  for this larger estimate is symbolized as  $n_1$  regardless of any system of subscripts that may have been used to designate the 2 groups. Thus the  $F$  that is used with the table is always unity or greater, even though the sampling distribution of  $F$  involves values less than unity. That is, if we were drawing successive samples from groups  $A$  and  $B$  and each time took  $F$  as  $s_a^2/s_b^2$ , regardless of which was the larger estimate, the sampling distribution of  $F$  would obviously involve values below unity as well as above unity. The table, however, is set up in terms of the greater-than-unity side of the sampling distribution.

If one wishes to judge whether 2 samples, either large or small, yield a difference in variability which is large enough to warrant concluding that the 2 population variabilities differ, he sets up the null hypothesis that no difference exists in the 2 population variances. Then, instead of dealing as usual with the difference between the 2 estimates, he takes their ratio. Obviously, the

$$* \quad y = \frac{\Gamma\left(\frac{n_1 + n_2}{2}\right) n_1^{n_1/2} n_2^{n_2/2}}{\Gamma\left(\frac{n_1}{2}\right) \Gamma\left(\frac{n_2}{2}\right)} \cdot \frac{F^{(n_1-2)/2}}{(n_1 F + n_2)^{(n_1+n_2)/2}}$$



departure of this ratio or  $F$  from unity reflects or depends upon the difference between the 2 variance estimates. If the value of  $F$ , computed with the larger estimate in the numerator, is so large that it is not reasonable to believe it a chance deviation from a true value of unity, the null hypothesis is rejected, and it is concluded that the 2 populations do not have the same variance. If  $F$  is small, i.e., near unity, the null hypothesis is accepted.

Now it happens that, although the  $F$  values given in Table F for the .05, the .01, and the .001 levels of significance hold for the major and very extensive uses of the  $F$  table to be discussed in Chapters 15, 16, and 17, these values are *not* applicable to the simple case where we wish to ascertain the probability of as great a difference (irrespective of direction, i.e., a hypothesis or decision requiring a two-tailed test) between the variances for 2 groups. For this particular case, an  $F$  which falls at, say, the .01 level signifies that as large a difference in *one* direction would occur 1 per cent of the time by chance. This is so because in placing the larger estimate in the numerator we are considering only 1 tail of the  $F$  distribution. In asking whether 2 variance estimates of, say, 10 and 25 based on 2 groups differ, i.e., lead to an  $F$  which departs significantly from unity (no difference), we should consider not only the probability of securing an  $F$  as large as  $25/10$  but also the probability of obtaining one as small as  $10/25$ . This, it will be observed, is exactly analogous to considering both positive and negative values for the  $z$  of formula (91) and then raising the question as to the probability of obtaining on a chance basis as large a difference, irrespective of direction. If we had this last probability, we would halve it to obtain the  $P$  for 1 direction only; conversely, if we had an  $F$  which fell at the  $P = .01$  level in the table, we would need to double .01 to secure the probability for as large a difference irrespective of direction. In other words, for this particular case, that of testing the significance between the variability for 2 groups, an  $F$  at the .01 point of the table means significance at the .02 level; an  $F$  at the .05 level means significance at the .10 level; and an  $F$  at the .001 level indicates significance at the .002 level. We will *not* have to make this type of adjustment when we come to the principal uses of  $F$  in connection with the analysis of variance.

For example, suppose that 50.21 and 147.62 are variance estimates available for 2 samples of 8 and 9 cases respectively.

The respective  $df$ 's would be 7 and 8. In computing  $F$  we have  $147.62/50.21 = 2.94$ , and  $n_1$  becomes 8, with  $n_2 = 7$ . Turning to Table F, we see that  $F$  would need to be 3.73 for the .05 level, which for this type of problem is the .10 level. Therefore the null hypothesis is not rejected. If we take the square roots of the 2 variance estimates, we get  $s$ 's of 7.09 and 12.15. By the  $F$  test, we are in effect saying that the difference between these 2  $s$ 's is not significant. As usual, this does not prove the null hypothesis—it becomes acceptable because we cannot with sufficient certainty reject it.

If the research hypothesis being tested or the decision to be made calls for a one-tailed test, the  $F$  values in Table F are applicable without further ado. As a matter of fact, if the null hypothesis is to be accepted unless  $s_a^2$  is significantly larger than  $s_b^2$ , one would not bother to compute  $F$  if  $s_a^2$  turned out to be smaller than  $s_b^2$ .

**Differences between several independent variances.** We have seen in the last chapter that  $\chi^2$  can be used to provide an over-all test of the difference between several independent proportions (p. 235) for  $C$  groups and also between  $C$  correlated proportions (p. 232). In the next chapter we shall see how an over-all test can be made for the differences between several means, either correlated or independent. We shall consider now an over-all test of the difference between 3 or more variance estimates. This test is not applicable when the variances are correlated (based on the same group or matched groups).

Suppose we have  $k$  variance estimates,  $s_1^2, s_2^2, \dots, s_i^2, \dots, s_k^2$ , based on  $m_1 - 1, m_2 - 1, \dots, m_i - 1, \dots, m_k - 1$  degrees of freedom respectively. Let  $N$  be the sum of the  $m$ 's. Compute the products: each  $s^2$  times its  $df$ . Sum these  $k$  products (the equivalent of summing the  $k$  sums of squares of deviations). Let  $s_w^2$  stand for this sum divided by  $N - k$ . Determine the log of each of the  $k$   $s^2$  values, then calculate the products: each  $\log s^2$  times the  $df$  for the given  $s^2$ . Sum these products, that is,  $\sum (m_i - 1) \log s_i^2$ , in which  $i$  takes on values from 1 to  $k$ . Determine the log of  $s_w^2$ , and compute

$$C = 1 + \frac{1}{3(k-1)} \left( \sum \frac{1}{m_i - 1} - \frac{1}{N - k} \right)$$

Finally, calculate the quantity

$$V = \frac{2.3026}{C} [(N - k) \log s_w^2 - \sum (m_i - 1) \log s_i^2] \quad (94)$$

The sampling distribution of  $V$  follows the  $\chi^2$  distribution with  $k - 1$  degrees of freedom. If  $V$  reaches the  $P = .05$  or  $P = .01$  or any a priori chosen level of significance, the differences between the  $k$  variances may be regarded as nonchance, hence the conclusion that the  $k$  groups have not been drawn from populations having equal variances. If  $V$  is not significant, one accepts the hypothesis that the groups have been drawn from populations having equal variances. The variances are said to be homogeneous. The procedure just described is known as Bartlett's test for the homogeneity of variances. It is appropriate for testing the assumption of homoscedasticity in bivariate correlation scattergrams.

## CHAPTER 15

### Analysis of Variance: Simple

The  $F$ , or variance, ratio defined in the previous chapter is applicable in a wide variety of situations. The general requirement is that we have 2 independent estimates of variance, which estimates are, on the basis of the null hypothesis, regarded as estimates of the same population value. If  $F$  is sufficiently large, the null hypothesis becomes suspect, and one draws a positive conclusion, the nature of which depends upon the given situation. Each application in this and the following chapter requires an *assumption of normality* and an *assumption of homogeneity of certain variances*; normality of what, and homogeneity of which variances, will need to be specified for each type of situation.

It will be recalled that under certain circumstances the squared correlation coefficient is interpretable in terms of the proportion of variance "explained." The idea is that variation can be broken down into component parts in such a way as to permit specification of the relative importance of the component sources. Back of this is the fact that variances are additive to a total variance, as shown when we derived formulas (37) and (37a), which are basic to the so-called variance theorem. Although this theorem is fundamental to the analysis of variance technique, it is not our aim to consider methods of estimating the proportion or percentage of variance due to a given source but rather to discuss ways of testing whether a possible source is contributing to the total variance to a statistically significant degree.

#### BREAKDOWN OF SUM OF SQUARES

Let us begin with the simple situation in which the total variation for a set of scores based on  $N$  individuals is possibly due in

part to the fact that the total group is heterogeneous with respect to some factor, such as socioeconomic level or age or racial origin or type of treatment or method used in memorizing or varying level of illumination—any factor which permits breaking down the total group into subgroups. In other words, the individuals or their scores can be classified into subgroups, or the total group can be regarded as made up of specified subgroups. For simplicity, let us assume that the subgroups are of the same size, say  $m$  cases per group, and that we have  $k$  groups. Let  $r$  stand for any subgroup; i.e.,  $r$  takes on values of 1, 2, 3,  $\dots$ ,  $k$ , and let the mean score for the groups be specified as  $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_r, \dots, \bar{X}_k$ , with  $\bar{X}$  as the mean for all groups combined (total mean). Although it is possible to use a precise notation, such as  $X_{ir}$ , to denote the score of any, the  $i$ th, person in group  $r$ , we shall in this chapter simply use  $X$  as the score for any individual.

We are now in a position to write an individual's score as a deviation from the total mean in terms of the deviation of his score from his group mean and the deviation of the group mean from the total mean. Thus, for a score in group  $r$ ,

$$(X - \bar{X}) = (X - \bar{X}_r) + (\bar{X}_r - \bar{X}) \quad (95)$$

which indicates 2 sources of variation: the variation of a group mean from the total mean and the variation of an individual's score from his group mean.

If we rewrite formula (95) specifically for group 1, we have

$$(X - \bar{X}) = (X - \bar{X}_1) + (\bar{X}_1 - \bar{X})$$

Squaring both sides gives

$$(X - \bar{X})^2 = (X - \bar{X}_1)^2 + (\bar{X}_1 - \bar{X})^2 + 2(\bar{X}_1 - \bar{X})(X - \bar{X}_1)$$

as the squared deviation, from the total mean, of any score in group 1. Each of the  $m$  persons in the group will have such a squared deviation score. We may indicate the sum of the squares for the  $m$  cases as

$$\Sigma(X - \bar{X})^2 = \Sigma(X - \bar{X}_1)^2 + \Sigma(\bar{X}_1 - \bar{X})^2 + 2(\bar{X}_1 - \bar{X})\Sigma(X - \bar{X}_1)$$

Note that in the last term the constants 2 and  $(\bar{X}_1 - \bar{X})$  have been taken from under the summation sign, and that  $\Sigma(X - \bar{X}_1)$ , being the sum of deviations of a set of scores about their own mean, will be exactly zero. Therefore, the last term vanishes.

Note also that the second right-hand term involves summing a constant, which is the same as multiplying it by the number of cases involved in the summation, i.e.,  $\Sigma(\bar{X}_1 - \bar{X})^2 = m(\bar{X}_1 - \bar{X})^2$ .

Thus we see that we may write the sum of squares (of deviations) for the first group and by analogy for the other groups as follows:

$$\text{1st group: } \Sigma(X - \bar{X})^2 = \Sigma(X - \bar{X}_1)^2 + m(\bar{X}_1 - \bar{X})^2$$

$$\text{2nd group: } \Sigma(X - \bar{X})^2 = \Sigma(X - \bar{X}_2)^2 + m(\bar{X}_2 - \bar{X})^2$$

$$\text{rth group: } \Sigma(X - \bar{X})^2 = \Sigma(X - \bar{X}_r)^2 + m(\bar{X}_r - \bar{X})^2$$

$$\text{kth group: } \Sigma(X - \bar{X})^2 = \Sigma(X - \bar{X}_k)^2 + m(\bar{X}_k - \bar{X})^2$$

If we summed the left-hand parts of the foregoing, we would obviously have the sum of squares of deviations for the entire set of  $N = km$  cases. This summing of sums, or double summation, can be conveniently indicated by using 2 summation signs, or  $\Sigma\Sigma(X - \bar{X})^2$ . We may sum the right-hand terms separately. The first term on the right involves summing sums, and the result can be indicated symbolically by  $\sum_r \Sigma(X - \bar{X}_r)^2$ , which implies that we first sum for each group, then sum over all groups. The first summation sign indicates that the subscript  $r$  takes in turn values running from 1 to  $k$ . The sum of the other right-hand terms can be written as  $m\sum_r (\bar{X}_r - \bar{X})^2$ .

Since adding of equations leads to an equation, we have

$$\Sigma\Sigma(X - \bar{X})^2 = \sum_r \Sigma(X - \bar{X}_r)^2 + m\sum_r (\bar{X}_r - \bar{X})^2 \quad (96)$$

as a means of expressing the fact that the total sum of squares (of deviations) can be broken down into 2 components, the first of which has to do with variation about group means, i.e., *within* groups, and the second of which involves variation of group means about the total mean, i.e., *between* groups. In other words, the total sum of squares is made up of 2 additive parts. If we divide both sides by  $N$  or  $km$ , we have the total variance broken into additive components, but for our present purposes we shall need unbiased estimates of variance, and hence it becomes necessary to divide through by degrees of freedom.

The correct *df* can be ascertained by examining the 3 sums of squares. For the total sum of squares we have 1 restriction,



the total mean, and as seen in Chapter 7, p. 106, the *df* will be  $N - 1$  or  $km - 1$ . The within-groups sum is based on  $N$  or  $km$  squares, but since these are about  $k$  different means there are  $k$  restrictions, or  $km - k (= N - k)$  degrees of freedom. The last or between-groups sum involves  $k$  means, varying more or less about the total mean; thus, aside from the  $m$  factor, it contains  $k$  squares with 1 restriction, and the *df* becomes  $k - 1$ . In other words, the  $k$  means are analogous to varying scores, and obviously the mean of these means will equal the total mean.

We may indicate the division of the 3 sums of squares by the proper *df*'s as follows:

$$\frac{\Sigma \Sigma (X - \bar{X})^2}{km - 1}, \quad \frac{\bar{\Sigma} \Sigma (X - \bar{X}_r)^2}{km - k}, \quad \frac{m \bar{\Sigma} (\bar{X}_r - \bar{X})^2}{k - 1}$$

Notice that we are no longer dealing with an equation. Why? Each division will result in a variance estimate, but these are not directly additive, which means that we cannot specify what proportion of the estimated total variance is due to the between-groups variation. The reader should note, however, that the *df*'s are additive:  $(km - 1) = (km - k) + (k - 1)$ .

Before examining the meaning of these 3 variance estimates, let us label them:  $s^2$  for the estimate of total variance,  $s^2_w$  for that based on the *within*-groups sum of squares,  $s^2_b$  for that based upon *between* groups. Variance estimates are sometimes referred to as "mean squares."

### MEANING OF VARIANCE ESTIMATES

In so far as one thinks of the total  $km$  cases as a sample drawn from 1 population,  $s^2$  will be the best unbiased estimate of the variance of the population,  $\sigma^2$ . If we think of the  $m$  cases for each of our  $k$  groups as samples from  $k$  possibly different populations, then  $s^2_w$  will be a composite estimate of the several population variances, a sort of average which makes sense if the population variances are equal; if the  $k$  groups have been drawn from just 1 population, this within-groups variance estimate or  $s^2_w$  will differ little from, but be somewhat smaller than,  $s^2$ . Note that  $s^2$  and  $s^2_w$  *cannot* be regarded as *independent* estimates because the 2 estimates are based on practically the same devia-

tions: extreme scores, in either direction, will tend to make both  $s^2$  and  $s^2_w$  large. If  $m$ , or the number of cases per group, is taken larger and larger and if the groups are regarded as belonging to the same population or populations differing in some respects but having the same mean and variance for the given trait or variate,  $s^2$  and  $s^2_w$  will tend to the same value,  $\sigma^2$ .

Let us next look at  $s^2_b$ . The division of  $m\sum(\bar{X}_r - \bar{X})^2$  by its  $df$  may be accomplished by dividing the sum factor by  $k - 1$ . In making this division we are dividing a sum of squares by degrees of freedom; hence the result will be a variance estimate. Let us use  $s^2_{x_b}$  as a symbol for this estimate. Then

$$s^2_b = \frac{m\sum(\bar{X}_r - \bar{X})^2}{k - 1} = ms^2_{x_b}$$

In order to understand the meaning of  $s^2_{x_b}$ , we may regard our  $k$  means as a sample of sample means from an indefinitely large supply of possible sample means for groups drawn from the same population. The variance for this universe of sample means is given by the standard error of mean formula, i.e.,  $\sigma^2_{x_b} = \sigma^2/m$ . If we were given the value of  $\sigma^2_{x_b}$  and told to determine the universe trait variance or  $\sigma^2$ , we would simply solve  $\sigma^2_{x_b} = \sigma^2/m$  for  $\sigma^2$ . Thus,  $\sigma^2 = m\sigma^2_{x_b}$ . If we had only an estimate of  $\sigma^2_{x_b}$ , such as  $s^2_{x_b}$ , we could use this estimate as a basis for estimating the trait variance; i.e.,  $ms^2_{x_b}$  can be taken as an estimate of  $\sigma^2$ . Since  $ms^2_{x_b} = s^2_b$ , we have  $s^2_b$  and  $s^2_w$  (see previous paragraph) as estimates of the same population variance.

These estimates should agree within the limits of chance, and being independent estimates of the same variance, the sampling distribution of their ratio is that of the  $F$  distribution. When an obtained  $F$  or  $s^2_b/s^2_w$  is larger than expected on the basis of chance sampling, the implication is that  $s^2_{x_b}$  is greater than expected by chance. How could this come about? Let us suppose that our  $k$  groups of  $m$  cases each have been drawn from  $k$  different populations, i.e., from populations with means which really differ. Under this circumstance the variation of the  $k$  sample means will spring from 2 sources. A part of the variance of the means will be due to sampling variation predictable by the formula for the standard error of the mean on the basis of  $m$  and the trait variance. A

second part of the variation in means will be due to the variation of the true (population) means of the  $k$  groups. If we let  $\sigma^2_{\bar{x}_h}$  represent the variance of obtained means and  $\sigma^2_{\bar{x}_\infty}$  the variance of the true group means, and if the several groups have the same population variance,  $\sigma^2$  (this is the assumption of homogeneity of variances), we should expect the following to hold exactly for an infinitely large number of groups and approximately for a small number of groups:  $\sigma^2_{\bar{x}_h} = \sigma^2 m + \sigma^2_{\bar{x}_\infty}$ . This is analogous to the commonly accepted expression used in connection with test reliability; namely, that the variance of obtained scores equals the variance of true scores plus error (of measurement) variance.

Multiplying the above by  $m$ , we have  $m\sigma^2_{\bar{x}_h} = \sigma^2 + m\sigma^2_{\bar{x}_\infty}$ . Thus, since  $m$  times the obtained variance of group means can be broken down into 2 components, it should be obvious that the estimate,  $m s^2_{\bar{x}_h}$ , may also be subject to 2 sources of variation.

In practice we don't have a priori knowledge of whether  $m\sigma^2_{\bar{x}_\infty}$  is or is not zero. What we have are 2 estimates of the population trait variance, that based on  $s^2_b$  (or  $m s^2_{\bar{x}_h}$ ) and that based on  $s^2_w$ . If the  $s^2_b$  estimate is significantly larger than  $s^2_w$ , i.e., if  $F$  or  $s^2_b/s^2_w$  is beyond the point for  $P = .01$  level of significance, it can be argued that  $s^2_b$  involves a source of variation over and above that of random sampling errors in the means, and hence that  $m\sigma^2_{\bar{x}_\infty}$  is real. This is, of course, equivalent to concluding that our  $m$  cases have been drawn from  $k$  groups with real differences in their population means.

Although the table of  $F$  requires that the larger of the 2 estimates be used as the numerator in computing the variance ratio, it should be noted that  $s^2_w$  cannot be significantly larger than  $s^2_b$  unless the operation of chance sampling has been restricted in some manner. In practical applications we are primarily and nearly always interested in the case in which  $s^2_b$  is the larger of the 2 estimates. If it is smaller than  $s^2_w$ , it is ordinarily not necessary to compute  $F$ .

We may now summarize the foregoing. When we have scores on  $k$  groups of  $m$  cases each, the total sum of squares can be broken down into 2 additive parts, that for between and that for within groups. Dividing by the appropriate degrees of freedom, the within sum of squares gives  $s^2_w$  as an estimate of the trait variance for the population, and  $s^2_b (= m s^2_{\bar{x}_h})$  yields a second and independent estimate of the same population variance. The

sampling variation of the ratio of these 2 estimates is that of the variance ratio,  $F$ , if the  $k$  groups belong to the same population. If  $s^2_b$  is significantly larger than  $s^2_w$ , which is an estimate of the population variance,  $s^2_b$  must be regarded as an estimate of the same variance *plus* variation due to real, nonchance, differences between the  $k$  groups.

If we let  $\rightarrow$  stand for "is an estimate of," then

$$s^2_w \rightarrow \sigma^2$$

$$s^2_b \rightarrow \sigma^2 + m\sigma^2_{\bar{x}_m}$$

The null hypothesis is that  $\sigma^2_{\bar{x}_m}$  is zero, and rejection of this hypothesis because  $s^2_b/s^2_w$  is significantly large implies that  $\sigma^2_{\bar{x}_m}$  is *not* zero, or that the  $k$  groups have not been drawn from the same population (or from populations with equal means). In other words, we have a technique that provides an over-all test for the significance of the differences between several means considered simultaneously.

For all the applications discussed in this chapter, it is *assumed* (1) that the  $m$  cases constituting each group have been drawn from a normally distributed population of scores for the trait or variable as measured and (2) that the  $k$  populations have the same variance. For large samples the first assumption can be checked by way of measures of skewness and kurtosis relative to their standard errors or by the chi square test of goodness of fit. Unfortunately neither of these checks is very sensitive for small samples. The second assumption may be evaluated, regardless of sample sizes, by Bartlett's test for the homogeneity of variances (p. 248). The reader will have noted that these 2 assumptions have to do with the distribution of scores within groups, which lead to the denominator,  $s^2_w$ , of  $F$ .

There is some evidence that moderate departure from normality and moderate lack of homogeneity of variances do not seriously disrupt the applicability of the technique. It is not easy to give a definition of "moderate," but it is known that violation of these assumptions leads to too many "significant"  $F$ 's. For example, an  $F$  which is apparently (from Table F) significant at the .05 level may really be significant only at the .07 level. One can guard against erroneously rejecting the null hy-

pothesis by choosing a more stringent level for judging significance.

**Computational formulas.** The required arithmetical labor can be shortened by resort to the general principle for computing the sum of squares of deviations inherent in formula (6a), p. 25:

$$\Sigma(X - \bar{X})^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N} = \frac{1}{N} [N\Sigma X^2 - (\Sigma X)^2]$$

Thus we would have

$$\Sigma\Sigma(X - \bar{X})^2 = \frac{1}{N} [N\Sigma\Sigma X^2 - (\Sigma\Sigma X)^2] \quad (97a)$$

for total sum of squares, in which the double summation indicates that the summing is over all groups. It can be shown by easy algebra that

$$\sum_r \Sigma(X - \bar{X}_r)^2 = \frac{1}{m} [m\Sigma\Sigma X^2 - \Sigma(\Sigma X)^2] \quad (97b)$$

for within sum of squares and that

$$m\sum_r (\bar{X}_r - \bar{X})^2 = \frac{1}{km} [k\Sigma(\Sigma X)^2 - (\Sigma\Sigma X)^2] \quad (97c)$$

for between sum of squares.

Accordingly, to compute the 3 sums of squares of deviations, we need to sum all the raw scores,  $\Sigma\Sigma X$ ; sum the squares of all the raw scores,  $\Sigma\Sigma X^2$ ; and sum the squares of the separate group sums,  $\Sigma(\Sigma X)^2$ . These sums can readily be obtained on a calculating machine by computing  $\Sigma X$  and  $\Sigma X^2$  separately for each group, squaring each  $\Sigma X$ , and then summing the several  $\Sigma X$  values for  $\Sigma\Sigma X$ , the  $\Sigma X^2$  values for  $\Sigma\Sigma X^2$ , and the  $(\Sigma X)^2$  values for  $\Sigma(\Sigma X)^2$ .

#### EXAMPLE: TESTING THE SIGNIFICANCE OF DIFFERENCES BETWEEN SEVERAL MEANS

To illustrate the application of the technique outlined above we shall use unpublished data of Wright\* on massed vs. dis-

\* Wright, Suzanne T., *Spacing of practice in verbal learning and the maturation hypothesis*, Unpublished Master's Thesis, Stanford University, California, 1946.

tributed practice in the learning of nonsense syllables by the anticipation method. The essential comparison is based on the amount of learning shown in 34 minutes by 5 ( $= k$ ) groups of 16 ( $= m$ ) cases each. The groups differed in length of rest intervals between trials and/or in the total number of trials, as indicated at the top of Table 36. The scores of all 80 subjects are

Table 36. NUMBER OF SYLLABLES CORRECTLY ANTICIPATED AT THE 34TH MINUTE OF PRACTICE

Group	1	2	3	4	5
Rest interval (minutes)	8	3.5	2	1.25	0
Number of trials	5	8	11	14	29
	5	8	9	11	17
	5	7	3	12	16
	1	4	9	15	18
	5	4	10	11	11
	8	7	5	10	15
	1	7	11	8	9
	2	5	9	13	18
	2	6	6	13	13
	2	8	7	5	12
	8	14	6	7	15
	4	8	16	11	8
	1	5	12	12	13
	3	1	11	12	7
	4	5	15	9	15
	4	8	13	16	15
	2	5	4	7	13
$m$	16	16	16	16	16
$\Sigma X$	87 +	102 +	146 +	172 +	215 = $\Sigma \Sigma X$ = 692
$\Sigma X^2$	279 +	708 +	1,550 +	1,982 +	3,050 = $\Sigma \Sigma X^2$ = 7,638
$(\Sigma X)^2$	3,249 +	10,404 +	21,316 +	29,584 +	46,225 = $\Sigma (\Sigma X)^2$ = 110,778
Means	3.56	6.38	9.12	10.75	13.44 $\bar{X}$ = 8.65

included in this table, and the necessary sums are given at the bottom of the table, separately for each group. Summing across yields the required double sums. The group means are also given, although not actually needed in determining  $F$ .

The sums of squares (of deviations) are obtained by substituting in formulas (97):

$$\Sigma \Sigma (X - \bar{X})^2 = \frac{1}{80} [80(7638) - (692)^2] = 1652.20$$

$$\Sigma \Sigma (X - \bar{X}_r)^2 = \frac{1}{16} [16(7638) - 110,778] = 714.38$$

$$m \Sigma (\bar{X}_r - \bar{X})^2 = \frac{1}{80} [5(110,778) - (692)^2] = 937.82$$



These sums of squares, along with the respective degrees of freedom and the resulting variance estimates are conveniently arranged in Table 37, usually referred to as a variance table. Note

Table 37. VARIANCE TABLE FOR DATA OF WRIGHT

Source	Sum of Squares	<i>df</i>	Variance Estimate
Between	937.82	4	$234.46 = s_b^2$
Within	714.38	75	$9.53 = s_w^2$
Total	1652.20	79	

that the sums of squares for between and within groups add to the sum for the total, which provides a check on the arithmetic involved in substituting in formulas (97). This does not check on the accuracy of the sums given in Table 36. Note also that the degrees of freedom add to the total *df*.

The variance ratio, or *F*, becomes  $234.46/9.53$  or  $24.60$ . With *df*'s of  $n_1 = 4$  and  $n_2 = 75$ , we refer to the table of *F* to learn whether  $24.60$  is larger than expected on the basis of chance. That this *F* is highly significant is immediately apparent when we note that for the given *df*'s an *F* of about  $5.2$  is significant at the .001 level. With the between-groups variance estimate significantly larger than that for within groups, we can conclude with high confidence that the 5 sets of scores have not been drawn from the same population of scores, or that amount of time spent in practice is a real source of variation. This is, of course, equivalent to saying that the several group means considered simultaneously differ significantly among themselves.

In the illustration just given the groups can be arranged in order before any of the data are seen, and additional credence can be placed in the results because the means follow this ordering. It should be understood, however, that the variance technique does not presuppose an a priori ordering of the several groups — it is generally applicable for testing the significance of the differences between group means regardless of prior considerations.

If one had available only the *CR* or *t* technique and wished to compare the means for 5 groups, it would ordinarily be necessary

to compute  $t$  or  $CR$  for each possible difference, and 5 means would lead to  $5 \times 4/2$  or 10 differences. Obviously, the variance method requires less computation, and furthermore it provides an over-all test of significance which is not subject to the fallacy inherent in singling out the comparison involving the largest obtained  $t$  or  $CR$ , a practice which is likely to capitalize on chance differences. After and only after it has been found that the over-all  $F$  is significant can one safely use the  $t$  technique to test the significance of the difference between any 2 of the group means. When we do this,  $s_w$  is used for the  $s$  required in the formula for  $t$ , p. 109. Thus, to check the significance of the difference between the means for groups 1 and 2 of the Wright data, we have

$$t = \frac{6.38 - 3.56}{\sqrt{\frac{9.53}{16} + \frac{9.53}{16}}} = \frac{2.82}{1.09} = 2.59$$

The variance estimate here used is based on 75 degrees of freedom; hence this  $t$  may be entered as a  $CR$  in the normal probability table. It is significant at the .01 level. Since group 1 differs still more from the remaining 3 groups, one would not bother to compute additional  $t$ 's for comparisons involving group 1. Actually the testing of the means for nonadjacent groups would scarcely be necessary, but note that, since the groups are of the same size, the  $t$  between any 2 means in Table 36 will involve the same denominator, 1.09, already used. The use of  $s_w^2$  as the  $s^2$  for the  $t$  test is logical in that  $s_w^2$  is based on all the available scores and hence is more dependable than an estimate based on just 2 groups.

#### SPECIAL CASE OF $F$ TEST WHEN $n_1 = 1$

If we had  $k = 2$  groups, the testing of the between-groups variance would appear to be much like testing the difference between 2 means. Let us examine this case by starting with the expressions for the sum of squares for 2 groups:

$$\text{1st group: } \Sigma(X - \bar{X})^2 = \Sigma(X - \bar{X}_1)^2 + m(\bar{X}_1 - \bar{X})^2$$

$$\text{2nd group: } \Sigma(X - \bar{X})^2 = \Sigma(X - \bar{X}_2)^2 + m(\bar{X}_2 - \bar{X})^2$$

Instead of using double summation signs, we may indicate the within-groups sum of squares as  $\Sigma(X - \bar{X}_1)^2 + \Sigma(X - \bar{X}_2)^2$ , and the between-groups sum of squares as  $m(\bar{X}_1 - \bar{X})^2$

+  $m(\bar{X}_2 - \bar{X})^2$ . The respective  $df$ 's will be  $2m - 2$  and 1. Indicating the division of the sums of squares by their  $df$ 's, we can write the variance ratio as

$$F = \frac{\frac{m(\bar{X}_1 - \bar{X})^2 + m(\bar{X}_2 - \bar{X})^2}{1}}{\frac{\Sigma(X - \bar{X}_1)^2 + \Sigma(X - \bar{X}_2)^2}{2m - 2}}$$

Since the number of cases for the 2 groups is the same, it is readily seen that the mean for 1 group will be exactly as far above the general mean ( $\bar{X}$ ) as the other group mean is below  $\bar{X}$ , or that  $\bar{X}$  will bisect the distance between  $\bar{X}_1$  and  $\bar{X}_2$ ; therefore  $(\bar{X}_1 - \bar{X})^2 = (\bar{X}_2 - \bar{X})^2 = \frac{1}{4}(\bar{X}_1 - \bar{X}_2)^2$ . The numerator for  $F$  becomes  $(m/2)(\bar{X}_1 - \bar{X}_2)^2$ . It will be noted that the denominator term, which defines  $s^2_x$ , is identical to the  $s^2$  defined on p. 110 in connection with the  $t$  test. Accordingly, we may write

$$F = \frac{\frac{m}{2}(\bar{X}_1 - \bar{X}_2)^2}{s^2}$$

Dividing both numerator and denominator by  $m/2$ , we have

$$F = \frac{(\bar{X}_1 - \bar{X}_2)^2}{s^2 \frac{2}{m}}$$

the square root of which is

$$\sqrt{F} = \frac{\bar{X}_1 - \bar{X}_2}{s \sqrt{\frac{2}{m}}} = \frac{\bar{X}_1 - \bar{X}_2}{s \sqrt{\frac{1}{m} + \frac{1}{m}}}$$

which is identical with a formula for  $t$ , p. 109; when  $k = 2$  or 2 groups are being compared,  $F = t^2$ . It can be shown that this is also true when the  $N$ 's or  $m$ 's for the 2 groups are unequal. In fact, it can be shown that, when  $n_1 = 1$ , the sampling distribution of  $F$  becomes the same as that for  $t^2$  providing the estimate based on between groups, i.e., that based on 1 degree of freedom, is

used as the numerator regardless of which of the 2 estimates is the larger. It is thus seen that the  $t$  test is a special case of the  $F$  test. Note that  $F$  involves the square of the difference between means; hence it provides a basis for judging whether a difference between means, irrespective of direction, is significant (cf. pp. 246-247). The  $CR$  technique for comparing the means of 2 large samples is also a special case of the more general  $F$  test. That is, when  $n_1 = 1$  and  $n_2$  is not small, the square root of  $F$  is  $CR$ , interpretable via the normal curve table (Table A of the Appendix).

### GROUPS OF UNEQUAL SIZE

When the number of cases varies from group to group, we may let  $m_1, m_2, \dots, m_r, \dots, m_k$  stand for the several  $N$ 's. The sum of squares for the  $r$ th group would be written as

$$\Sigma(X - \bar{X})^2 = \Sigma(X - \bar{X}_r)^2 + m_r(\bar{X}_r - \bar{X})^2$$

and the double summation over all groups would be

$$\Sigma\Sigma(X - \bar{X})^2 = \Sigma\Sigma(X - \bar{X}_r)^2 + \Sigma m_r(\bar{X}_r - \bar{X})^2$$

which differs from formula (96) in that the varying  $m$ 's must be left under the summation sign in the last term. In specifying the degrees of freedom, we must replace  $km$  by  $N$ , where  $N$  is the total cases for all groups. The respective  $df$ 's become  $N - 1$ ,  $N - k$ , and  $k - 1$ . The computational formulas are changed to

$$\Sigma\Sigma(X - \bar{X})^2 = \Sigma\Sigma X^2 - \frac{(\Sigma\Sigma X)^2}{N} \quad \text{for total sum} \quad (98a)$$

$$\Sigma\Sigma(X - \bar{X}_r)^2 = \Sigma\Sigma X^2 - \Sigma \frac{(\Sigma X)^2}{m_r} \quad \text{for within sum} \quad (98b)$$

$$\Sigma m_r(\bar{X}_r - \bar{X})^2 = \Sigma \frac{(\Sigma X)^2}{m_r} - \frac{(\Sigma\Sigma X)^2}{N} \quad \text{for between sum} \quad (98c)$$

Note that the second term for the within sum (and the first for the between) requires that for each group the square of the sum of its scores be first divided by its  $m$ ; then the several quotients are summed. An additional row would be needed along the bot-

tom of Table 36 for these quotients if the  $m$ 's differed, or one might replace the  $(\Sigma X)^2$  row by  $(\Sigma X)^2/m_r$  values.

A variance table (like Table 37) may be formed, and  $F$  taken to equal  $s^2_b/s^2_w$  as before. The same interpretation holds: if  $F$  is significantly large, i.e., if  $s^2_b$  is significantly larger than  $s^2_w$ , the variation of the several group means among themselves is larger than expected on the basis of sampling; hence nonchance differences exist between the groups. The student who attempts, for the situation of unequal  $m$ 's, to reorient the logic leading to the idea that  $s^2_w$  is an estimate of  $\sigma^2$  and that  $s^2_b$  is an estimate of  $\sigma^2$  plus a possible  $m\sigma^2_{\bar{x}_w}$  will encounter some difficulty. Suffice it to say here that, if  $s^2_b$  is significantly larger than  $s^2_w$ , it can be concluded that the component involving the variance  $\sigma^2_{\bar{x}_w}$  is not zero. That is, when the groups have been drawn from populations having different means,  $s^2_b$  may be larger than  $s^2_w$  because of this additional source of variation even though it is not easy to regard this variation in terms of  $\sigma^2_{\bar{x}_w}$  times a varying  $m$ .

Thus the  $F$  technique may be applied as a test of the significance of the difference between 2 or more means based on large or small samples of equal or unequal size (per group) regardless of whether there is an a priori basis for arranging the groups in order. It might be said parenthetically that the scientific hypothesis being tested will specify the direction of differences if such are expected.

### TESTING THE SIGNIFICANCE OF THE CORRELATION RATIO

If the definitions of the correlation ratio,  $\eta$  (pp. 207-208), are reexamined, it is readily seen that for 1 variable the within-arrays variance is the same as the within-groups variance, the grouping being made on the basis of intervals on another variable. Also the variance of array means is the same as between-groups variance. We recall, however, that the correlation ratio, as defined, does not involve the idea of variance estimates. It should be rather obvious that, unless the between-arrays (groups) variance is significantly larger than expected on the basis of sampling errors in the array means, a correlation ratio cannot be deemed significant.

For purposes of exposition we shall outline the procedure for testing the significance of  $\eta_{yx}$ , for which we shall use the simpler symbol  $\eta$ . The grouping will be on the basis of the intervals on the  $X$  variable, and the required sum of squares will be in terms of  $Y$ . The sums of squares and their respective degrees of freedom will be

$$\sum\sum_{(N-1)} (Y - \bar{Y})^2 = \sum\sum_{(N-k)}^r (Y - \bar{Y}_r)^2 + \sum_{(k-1)}^r m_r (\bar{Y}_r - \bar{Y})^2$$

for  $k$  arrays with varying number,  $m_r$ , of cases per array. From the definition formula of the correlation ratio, we have

$$\eta^2 = 1 - \frac{\sigma_{ay}^2}{\sigma_y^2}$$

which becomes, in the notation of this chapter,

$$\eta^2 = 1 - \frac{\sum\sum^r (Y - \bar{Y}_r)^2 / N}{\sum\sum (Y - \bar{Y})^2 / N}$$

Since  $N$  cancels, we see that the following holds:

$$\begin{aligned} \sum\sum^r (Y - \bar{Y}_r)^2 &= (1 - \eta^2) \sum\sum (Y - \bar{Y})^2 \\ &= \text{within sum of squares} \quad (99) \end{aligned}$$

From the alternate expression for  $\eta$  we have

$$\eta^2 = \frac{\sigma_{m_y}^2}{\sigma_y^2}$$

which becomes

$$\eta^2 = \frac{\sum m_r (\bar{Y}_r - \bar{Y})^2 / N}{\sum\sum (Y - \bar{Y})^2 / N}$$

which leads to

$$\begin{aligned} \sum m_r (\bar{Y}_r - \bar{Y})^2 &= \eta^2 \sum\sum (Y - \bar{Y})^2 \\ &= \text{between sum of squares} \quad (100) \end{aligned}$$

When we wish to divide the sum of squares of formula (99) or (100) by the proper  $df$ , we may choose either the left- or right-hand part as representing the sum of squares. Thus the between-



arrays estimate may be written as

$$s_b^2 = \frac{\eta^2 \Sigma \Sigma (Y - \bar{Y})^2}{k - 1}$$

and that for within arrays as

$$s_w^2 = \frac{(1 - \eta^2) \Sigma \Sigma (Y - \bar{Y})^2}{N - k}$$

The ratio,  $F = s_b^2/s_w^2$ , may be written as

$$\begin{aligned} F &= \frac{\eta^2 \Sigma \Sigma (Y - \bar{Y})^2 / (k - 1)}{(1 - \eta^2) \Sigma \Sigma (Y - \bar{Y})^2 / (N - k)} \\ &= \frac{\eta^2 / (k - 1)}{(1 - \eta^2) / (N - k)} \end{aligned}$$

It is accordingly seen that for fixed  $df$ 's the value of  $F$ , even though computed from the sums rather than from their equivalents in terms of  $\eta^2$ , can be thought of as depending upon the size of  $\eta^2$ ; therefore a significant  $F$  indicates a significant correlation ratio.

With the 3 sums of squares computed, we can readily determine whether any correlation in the sense of the correlation ratio exists, and we also have the necessary sums for calculating  $\eta$  if it is desired to have this measure of the degree of correlation. A significant  $F$  does not, however, mean a high correlation ratio; with  $N$  large, a low  $\eta$  can possess statistical significance.

The computation of the sums of squares is accomplished by means of formulas (98) with the  $X$ 's replaced by  $Y$ 's.

### SIGNIFICANCE OF LINEAR CORRELATION

An appreciable correlation between 2 variables which are linearly related implies that the slopes of the regression lines are not zero, which in turn implies that the variance of predicted values is large enough to have some kind of statistical significance. The variance technique may be used as a test of the significance of linear regression.

Suppose that we develop the argument in terms of the regression of  $Y$  on  $X$ . We may write the linear equation for predicting  $Y$  from  $X$  as  $Y' = BX + A$ . If we think of this regression line

as having been drawn on the scatter diagram, it can readily be seen that the deviation of any person's  $Y$  value from the mean of the  $Y$ 's can be expressed in terms of its deviation from the regression line (or predicted value) plus the deviation of the predicted value from the mean of the  $Y$ 's:

$$(Y - \bar{Y}) = (Y - Y') + (Y' - \bar{Y})$$

in which  $Y'$  will vary from person to person in accordance with his  $X$  score. If we square all such  $(Y - \bar{Y})$  deviations and sum over all cases, we get

$$\begin{aligned}\Sigma\Sigma(Y - \bar{Y})^2 \\ &= \Sigma[(Y - Y') + (Y' - \bar{Y})]^2 \\ &= \Sigma(Y - Y')^2 + \Sigma(Y' - \bar{Y})^2 + 2\Sigma(Y - Y')(Y' - \bar{Y})\end{aligned}$$

for which double summation signs are not needed for clarity even though the summing is over all cases. The last or cross-product term has to do with a possible relationship between predicted values and residuals, but, as was shown in Chapter 9, this correlation is always zero, and hence this last term vanishes.

Therefore the sum of squares can be broken down into 2 components: residuals or within arrays about the regression line and a part depending on the variation of the predicted values about the mean. If the correlation between  $X$  and  $Y$  were zero, this latter component would be zero because one would predict  $\bar{Y}$  for all cases. The departure of this sum of squares or of a variance estimate based thereon from zero might lead one to conclude that real correlation exists in the population being sampled if it were not for the fact that sampling errors ordinarily operate so as to prevent the obtaining of zero correlation.

Before attempting to understand the operation of chance sampling, we should consider the degrees of freedom associated with the sums of squares. As usual, the total sum of squares is based on  $N - 1$  degrees of freedom. The  $df$  for  $\Sigma(Y - Y')^2$  may not be immediately obvious, but note that, if  $N = 2$  and variation exists for both  $X$  and  $Y$ , the regression line would necessarily pass through the 2 points defined by the pair of scores,  $r$  would be unity, and  $\Sigma(Y - Y')^2$  would be zero. In other words, with  $N = 2$ , there is no freedom for deviation from the regression line. From this it would be inferred that  $N$  needs to be reduced by 2,

or that  $df = N - 2$ , a deduction which is consistent with the fact that, in fitting a straight line, 2 constants are determined from the data, and hence 2 restrictions are imposed on the  $N$  deviations of the type  $(Y - Y')$ .

Since the  $df$ 's for the component sums of squares are additive to that for the total, one can determine the  $df$  for the regression or  $\Sigma(Y' - \bar{Y})^2$  term by subtracting the  $df$  for residuals from that for the total:  $(N - 1) - (N - 2) = 1$  as the  $df$  for the regression term. But determination of a  $df$  by subtraction does not permit the additive check on the correctness of the  $df$ 's which is possible in case each  $df$  is ascertained separately on the basis of some principle. By what principle could one determine that for the regression sum of squares the proper  $df$  is 1? The value of  $\Sigma(Y' - \bar{Y})^2$  will not be changed by shifting from gross scores to deviation scores, i.e., by moving the origin to the intersection of  $\bar{X}$  and  $\bar{Y}$ . It will be recalled that the regression equation in deviation units is  $y' = bx$  (where  $b = B$  of the gross score form), and accordingly we may write

$$\Sigma(Y' - \bar{Y})^2 = \Sigma(y' - \bar{y})^2 = \Sigma(y' - 0)^2 = \Sigma(bx)^2 = b^2 \Sigma x^2$$

which permits us to examine the source or sources of variation in the regression sum of squares. Its value depends upon  $b^2$  and  $\Sigma x^2$ , but the value of  $\Sigma x^2$  does not depend upon the degree of correlation. For a fixed set of  $X$ 's, the freedom of  $\Sigma(Y' - \bar{Y})^2$  to vary springs from  $b$ , i.e., from *one* value; therefore the  $df$  is 1. A slightly different way of considering the question is to note that, since  $b = r(\sigma_y/\sigma_x)$  and  $\Sigma x^2 = N\sigma_x^2$ , the sum of the squares of the predicted values can be written as

$$\Sigma(Y' - \bar{Y})^2 = r^2 \frac{\sigma_y^2}{\sigma_x^2} (N\sigma_x^2) = Nr^2\sigma_y^2$$

from which it can be argued that, since the variation in predicted values is a function neither of  $N$  nor of the variance of the trait being predicted, it is a function of *one* value, the degree of correlation.

Now let us return to a brief consideration of sampling or of the meaning of the variance estimates which result from dividing the

sums of squares by their  $df$ 's. On the basis of the null hypothesis, that the degree of linear correlation is zero for the population being sampled, the regression line for the population would pass through  $\bar{Y}$ , with zero slope or parallel to the  $x$  axis. Hence  $(Y - Y')$  will equal  $(Y - \bar{Y})$ , and the variance of the residuals will equal the total variance of the  $Y$ 's. A sample from the population will seldom yield zero correlation (or zero regression), and therefore the residuals will tend to be somewhat reduced, or  $\Sigma(Y - Y')^2$  will tend to be less than  $\Sigma(Y - \bar{Y})^2$ . It can be shown that  $\Sigma(Y - Y')^2/(N - 2)$  gives an unbiased estimate of the population variance when no correlation exists in the population.

That the estimate based on the regression sum of squares,  $\Sigma(Y' - \bar{Y})^2$ , divided by  $df = 1$ , is also an unbiased estimate of the same population variance may not seem plausible, nor is it easily explained in an elementary treatment. For any sample,  $\Sigma(Y' - \bar{Y})^2$  equals the difference between  $\Sigma(Y - \bar{Y})^2$  and  $\Sigma(Y - Y')^2$ , and it can be demonstrated that on the average the value of  $\Sigma(Y - \bar{Y})^2 - \Sigma(Y - Y')^2$  will equal  $\Sigma(Y - \bar{Y})^2/(N - 1)$ , or that the mean value of  $\Sigma(Y' - \bar{Y})^2$  for successive samples will be  $\Sigma(Y - \bar{Y})^2/(N - 1)$ . Since the latter is an unbiased estimate of the population variance, it follows that  $\Sigma(Y' - \bar{Y})^2/1$  must be an estimate of the same variance.

Of the 3 variance estimates, only the estimates based on residuals and on regression are independent. The sampling distribution of their ratio is that of  $F$ . Let  $s_r^2$  stand for the estimate based on the residual sum of squares and  $s_p^2$  stand for the estimate based on predictions by a linear regression function. Then, if  $s_p^2/s_r^2$ , with  $n_1 = 1$  and  $n_2 = N - 2$ , falls at or beyond the .01 level of significance, the null hypothesis becomes suspect. This means that the  $s_p^2$  estimate is larger than expected on the basis of sampling, from which it may be inferred that regression is a real source of variation in  $\Sigma(Y' - \bar{Y})^2$ , i.e., that the slope of the regression for the population is not zero, or that some correlation exists.

We have already noted that

$$\Sigma(Y' - \bar{Y})^2 = Nr^2\sigma_y^2$$

Since  $\Sigma(Y - Y')^2$  divided by  $N$  equals the error of estimate vari-

ance, previously proved to equal  $\sigma_y^2(1 - r^2)$ , it follows readily that

$$\Sigma(Y - Y')^2 = N(1 - r^2)\sigma_y^2$$

Accordingly

$$s_p^2 = \frac{\Sigma(Y' - \bar{Y})^2}{1} = \frac{Nr^2\sigma_y^2}{1}$$

and

$$s_r^2 = \frac{\Sigma(Y - Y')^2}{N - 2} = \frac{N(1 - r^2)\sigma_y^2}{N - 2}$$

Therefore

$$F = \frac{Nr^2\sigma_y^2/1}{N(1 - r^2)\sigma_y^2/(N - 2)} = \frac{r^2}{(1 - r^2)/(N - 2)}$$

which is the square of the  $t$  given earlier (p. 146) for testing the significance of  $r$ . Thus, again we have  $F = t^2$ , when  $n_1 = 1$ .

The reader will have noted that, since the required sums of squares and the resulting  $F$  can readily be expressed in terms of  $r$ , there is no need to worry further about a computational scheme for securing the sums of squares. The easier thing to do is simply to compute  $r$ . After that is done, either the  $F$  or the  $t$  test may be used for judging whether the correlation is significant. This discussion of the linear correlation problem here should help the student appreciate the generality of the analysis of variance technique and should also provide him with relevant concepts for understanding the test for curvilinearity of regression, to which we now turn.

### TESTING LINEARITY OF REGRESSION

We have seen that the correlation ratio is a general measure of the degree of correlation and that  $r$  measures the degree of linear relationship. Even though the regression of  $Y$  on  $X$  for a population be exactly linear, it will be found for a sample that the means of the arrays will show some deviation from a straight line; hence, as previously pointed out, the correlation ratio will tend to be larger than  $r$ . How large should the difference between  $\eta$  and  $r$  be before one suspects nonlinearity, or how much can the array means deviate from a straight line by chance? Before the development of the analysis of variance technique, the inadequate Blakeman criterion was used to answer the foregoing. In presenting the

currently accepted method, we shall carry the argument through on the basis of the regression of  $Y$  on  $X$ .

Imagine a scatter diagram with regression line drawn and the array mean located in each vertical array. For a score in the  $r$ th array, the deviation of  $Y$  from  $\bar{Y}$  can be thought of in terms of its deviation from the array mean,  $\bar{Y}_r$ , plus the deviation of the array mean from the predicted value,  $Y'_r$ , plus the deviation of the predicted value from the total mean. In symbols,

$$(Y - \bar{Y}) = (Y - \bar{Y}_r) + (\bar{Y}_r - Y'_r) + (Y'_r - \bar{Y})$$

Squaring and summing for the  $m_r$  cases in each array and then summing over all  $k$  arrays (equivalent to summing over all groups), we have

$$\begin{aligned} \Sigma \Sigma (Y - \bar{Y})^2 \\ = \Sigma \Sigma (Y - \bar{Y}_r)^2 + \Sigma m_r (\bar{Y}_r - Y'_r)^2 + \Sigma m_r (Y'_r - \bar{Y})^2 \end{aligned}$$

the cross-product terms having vanished because the component parts are uncorrelated.

The first component is a sum of squares based on within-array variation with  $N - k$  degrees of freedom. We encountered this in checking the significance of the correlation ratio, and we then labeled as  $s^2_w$  the variance estimate based thereon.

The second sum involves deviations of array means from linear regression. Its  $df$  will be  $k - 2$  since there are  $k$  means and 2 restrictive constants in  $Y'_r$ . If  $k = 2$ , the 2 means cannot vary from the fitted line. Let us use  $s^2_d$  as a symbol for the variance estimate based on this sum of squares.

The third sum, which has to do with the part of the total variance predictable by means of linear regression, is very similar to that occurring a few pages earlier in connection with the  $F$  test of the correlation coefficient. It differs only in that the same value is predicted for all cases within an array regardless of their location in the  $X$  interval defining the array. This is equivalent to a linear prediction of the mean of the array. Actually, the numerical value of  $\Sigma (Y' - \bar{Y})^2$  as calculated by  $Nr^2\sigma_y^2$ , which equals  $r^2 \Sigma \Sigma (Y - \bar{Y})^2$ , will be the same as  $\Sigma m_r (Y'_r - \bar{Y})^2$  computed directly, provided  $r$  was originally determined from a scatter diagram with the same intervals now being used to define the arrays. We have already seen that the  $df$  for this sum is 1, and we have used  $s^2_p$  as a symbol for the estimate based thereon.



It will be recalled that, in the scheme for testing the significance of the correlation ratio, the total sum of squares was broken down into a within-array and a between-array part. We now have a breakdown into within array (as before) plus 2 additional parts — the sum  $\sum_r (\bar{Y}_r - \bar{Y})^2$  is broken into

$$\sum_r (\bar{Y}_r - Y'_r)^2 + \sum_r (Y'_r - \bar{Y})^2$$

It will also be recalled that

$$\sum_r (\bar{Y}_r - \bar{Y})^2 = \eta^2 \Sigma \Sigma (Y - \bar{Y})^2$$

and that

$$\sum_r (Y'_r - \bar{Y})^2 = r^2 \Sigma \Sigma (Y - \bar{Y})^2$$

By subtraction, we see that the new sum,  $\sum_r (\bar{Y}_r - Y'_r)^2$ , is equivalent to  $(\eta^2 - r^2) \Sigma \Sigma (Y - \bar{Y})^2$ .

For convenience, we shall now assemble in an analysis of variance table the several symbolic expressions having to do with testing the significance of (1) the correlation ratio, (2) the linear regression coefficient, and (3) nonlinearity of regression. Table 38

Table 38. ANALYSIS OF VARIANCE FUNCTIONS FOR BIVARIATE CORRELATION

Source of Variation	Sum of Squares	Equivalent	df	Estimate
(a) Linear regression	$\sum_r (Y'_r - \bar{Y})^2$	$= r^2 \Sigma \Sigma (Y - \bar{Y})^2$	1	$s_p^2$
(b) Deviation of means from line	$\sum_r (\bar{Y}_r - Y'_r)^2$	$= (\eta^2 - r^2) \Sigma \Sigma (Y - \bar{Y})^2$	$k - 2$	$s_d^2$
(c) Between-array means	$\sum_r (\bar{Y}_r - \bar{Y})^2$	$= \eta^2 \Sigma \Sigma (Y - \bar{Y})^2$	$k - 1$	$s_b^2$
(d) Within arrays	$\sum_r \Sigma (Y - \bar{Y}_r)^2$	$= (1 - \eta^2) \Sigma \Sigma (Y - \bar{Y})^2$	$N - k$	$s_w^2$
(e) Residual from line	$\sum_r \Sigma (Y - Y'_r)^2$	$= (1 - r^2) \Sigma \Sigma (Y - \bar{Y})^2$	$N - 2$	$s_r^2$
(f) Total	$\Sigma \Sigma (Y - \bar{Y})^2$		$N - 1$	

gives the sources of variation, the sums of squares and their equivalents in terms of  $r$  or  $\eta$ , the degrees of freedom, and a symbol for each of the variance estimates. Note, in review, that for the sums of squares, their equivalents, and the  $df$ 's, the following additions hold true:

$$(a) + (b) = (c)$$

$$(a) + (e) = (f)$$

$$(c) + (d) = (f)$$

$$(a) + (b) + (d) = (f)$$

The several useful and permissible  $F$ 's, or ratios of independent and unbiased variance estimates, along with the proper  $df$ 's ( $n_1$  and  $n_2$  values) for entering the table of  $F$ , may be stated in summary form:

$$F_1 = s_b^2/s_w^2; \quad n_1 = k - 1, \quad n_2 = N - k: \quad \text{significance of correlation ratio}$$

$$F_2 = s_p^2/s_r^2; \quad n_1 = 1, \quad n_2 = N - 2: \quad \text{significance of linear correlation}$$

$$F_3 = s_d^2/s_w^2; \quad n_1 = k - 2, \quad n_2 = N - k: \quad \text{significance of curvilinearity}$$

We have already discussed the first 2 of these  $F$ 's. If we write the third in terms of sums and  $df$ 's, we have

$$\begin{aligned} F_3 &= \frac{s_d^2}{s_w^2} = \frac{\sum m_r (\bar{Y}_r - Y'_r)^2 / (k - 2)}{\sum \Sigma (Y - \bar{Y}_r)^2 / (N - k)} \\ &= \frac{(\eta^2 - r^2) \Sigma \Sigma (Y - \bar{Y})^2 / (k - 2)}{(1 - \eta^2) \Sigma \Sigma (Y - \bar{Y})^2 / (N - k)} \\ &= \frac{(\eta^2 - r^2) / (k - 2)}{(1 - \eta^2) / (N - k)} \end{aligned}$$

which indicates definitely that its value, for given  $df$ 's, is a reflection of the difference between the correlation ratio and the correlation coefficient. Therefore, in testing the significance of the

variation of array means from linear regression, we are testing the significance of the difference between  $\eta$  and  $r$ . If  $F_3$  falls beyond the .01 probability level, the hypothesis of linear regression for the population being sampled is rejected. When this happens, it follows that the correlation coefficient and a linear regression function for  $Y$  on  $X$  are not appropriate measures to use in describing the relationship.

If one is also interested in testing the significance of the correlation ratio for  $X$  on  $Y$  and the linearity of the horizontal array means, the analysis is carried through with  $X$ 's substituted for  $Y$ 's. Since the number of grouping intervals on the 2 axes need not be the same, the value of  $k$  may differ for the 2 analyses.

### ILLUSTRATIVE PROBLEM: $r$ , $\eta$ , AND CURVILINEARITY

The foregoing 3 tests of significance and the computations necessary thereto may be illustrated by the data of Table 39,

Table 39. BIVARIATE SCATTER FOR INITIAL AND FINAL SCORES OF 92 BOYS ON KOERTH PURSUIT ROTOR

Y = Final Score      Code		X = Initial Score							$f_y$	
		0	30	60	90	120	150	180		210
740	11				1					1
700	10		1	2	1	1		2	2	9
660	9	1	1	1	4	3		1	2	13
620	8	2	8	2	2	2		1		17
580	7	3	3	7	1	1			1	16
540	6	2	8	5						15
500	5	2	5	3	1					11
460	4	3	1							4
420	3	2								2
380	2									
340	1	3								3
300	0	1								1
$f_x = m_r$		19	27	20	10	7	0	4	5	92 = N
$\Sigma Y$		89	181	139	85	60	0	37	45	636
$\Sigma Y^2$		547	1269	1007	747	520	0	345	411	4846
$(\Sigma Y)^2/m_r$		416.89	1213.37	966.05	722.50	514.29	0	342.25	405.00	4580.35

which gives the bivariate distribution for the relationship between initial (sum of scores on trials 1-4) and final (trials 67-70) performance on the Koerth pursuit rotor. Since it is logical to be concerned with the prediction of final from initial score, or the regression of  $Y$  on  $X$ , we shall be dealing with variations on the  $Y$  variable.

In the first place, the correlation coefficient is computed from the scatter diagram by the method given in Chapter 8. Its value of .5687 is about .01 lower than the coefficient computed from a scatter with twice as many intervals. The use of so few intervals for the  $X$  variable would obviously not be recommended for the computation of  $r$ , but in this illustration it is convenient because of page-space limitations. There is the additional consideration that for computing the correlation ratio one should avoid having too few cases per array, which if the sample is small may mean only a few intervals on the independent variable. At least 12 intervals should be used for the dependent variable. In checking on linearity, it is necessary that we calculate  $r$  from a scatter with the same grouping intervals used in computing  $\eta$ , and no corrections for grouping error are needed.

For the computation of the correlation ratio and for the testing of its significance, we need the within arrays, the between arrays, and the total sum of squares. These may be computed from coded scores (deviations from an arbitrary origin in terms of step intervals), and the entire analysis may be carried through on the basis of coded scores, so that clumsily large figures are avoided. The reader who wishes to follow the computational procedure will need to note the following features of Table 39. The marginal frequencies on the right are for all the  $Y$  scores, and the  $f_r$ 's along the bottom margin are the  $m_r$ 's, or cases per array. For each vertical array and for the right-hand margin,  $\Sigma Y$  and  $\Sigma Y^2$  are computed in terms of coded values (these correspond to  $\Sigma d$  and  $\Sigma d^2$  of Chapter 3). Summing across the  $\Sigma Y$  and  $\Sigma Y^2$  rows should yield the  $\Sigma Y$  and  $\Sigma Y^2$  obtained from the marginal distribution. For this problem,  $\Sigma \Sigma Y = 636$  and  $\Sigma \Sigma Y^2 = 4846$ . The last row, containing the several values of  $(\Sigma Y)^2/m_r$ , is summed across for the needed  $\sum_r \frac{(\Sigma Y)^2}{m_r}$ , which is 4580.35 in this example. There is no check on this figure by calculations based on the margin.

In order to get the sums of squares of deviations, the values 636, 4846, and 4580.35 are substituted in formulas (98) with  $X$  replaced by  $Y$ .

$$\Sigma(Y - \bar{Y})^2 = 4846 - \frac{636^2}{92} = 449.30$$

$$\sum_r (Y - \bar{Y}_r)^2 = 4846 - 4580.35 = 265.65$$

$$\sum_r (\bar{Y}_r - \bar{Y})^2 = 4580.35 - \frac{636^2}{92} = 183.65$$

By formula (100) we now obtain

$$\eta^2 = \frac{183.65}{449.30} = .40874; \quad \eta = .639$$

which is the correlation ratio for  $Y$  on  $X$ .

The other sums of squares called for in schematic Table 38 may be calculated from their equivalents in terms of  $r^2$  and/or  $\eta^2$ . Note that  $r^2 = .5687^2 = .32342$ .

$$\sum_r (Y'_r - \bar{Y})^2 = (.32342)(449.30) = 145.31$$

$$\sum (Y - Y'_r)^2 = (1 - .32342)(449.30) = 303.99$$

$$\sum_r (\bar{Y}_r - Y'_r)^2 = (.40874 - .32342)(449.30) = 38.34$$

The several sums of squares and their respective degrees of freedom are set forth in Table 40, which contains also the variance

Table 40. ANALYSIS OF VARIANCE TABLE FOR REGRESSION OF FINAL ( $Y$ ) ON INITIAL SCORE FOR DATA OF TABLE 39

Source	Sum of Squares	df	Variance Estimate
Linear regression	145.31	1	$145.31 = s_p^2$
Deviation of means from line	38.34	5	$7.67 = s_d^2$
Between-array means	183.65	6	$30.61 = s_b^2$
Within arrays	265.65	85	$3.13 = s_w^2$
Residual from line	303.99	90	$3.38 = s_r^2$
Total	449.30	91	

estimates obtained by dividing the sums of squares by their  $df$ 's. From these variance estimates, we have the following:

For testing the significance of the correlation ratio we have  $F_1 = 30.61/3.13 = 9.8$ , which for  $n_1 = 6$  and  $n_2 = 85$  is highly significant. The .001 level of significance requires an  $F$  of about 4.0.

For testing the significance of linear correlation, i.e.,  $r$ , we have  $F_2 = 145.31/3.38 = 43.0$ , which for  $n_1 = 1$  and  $n_2 = 90$  is likewise highly significant, the .001 level being at an  $F$  of about 11.6.

For testing linearity of regression, i.e., the departure of the array means from a straight line, we have  $F_3 = 7.67/3.13 = 2.5$ , which for  $n_1 = 5$  and  $n_2 = 85$  is near the .05 level of significance. Thus the apparent departure from linearity in Table 39 is not sufficiently great to lead to rejection of the hypothesis of linearity; one would, however, question the hypothesis. This is an example of borderline significance which calls for drawing another sample or adding more cases before one sets forth a conclusion. For the problem at hand, a second sample of 90 boys yields a scatter diagram much like that of Table 39, so we would reject the hypothesis of linearity of regression.

The student should keep in mind that the test for linearity can lead to the definite conclusion that the regression is curvilinear (if  $F$  is large enough), whereas a low  $F$  does not prove linearity. Why?

If the hypothesis of linearity is disproved, it follows that the correlation coefficient is not a suitable figure for describing the relationship. The correlation ratio can be used to describe the *degree* of association, but the *form* of the relationship should be described by a fitted curve or by a verbal description of the general curve tendency of the array means. Some readers will have noted that the correlation ratio cannot be considered very descriptive of the data of Table 39 because of heteroscedasticity. As a matter of fact, the lack of homoscedasticity may also mean that our analysis of variance test for linearity is subject to question in that the assumption of homogeneity of variance is violated. The possible extent and direction of the error due to this failure of the groups, as defined by intervals on the  $x$  axis, to exhibit like variances cannot be specified, but it is doubtful whether the error is serious.



## APPLICATION TO MULTIPLE CORRELATION

The reader may recall that the methods given in Chapter 11 for judging the significance of the multiple correlation coefficient involved unsatisfactory approximations. In so far as we are interested in testing the deviation of a multiple  $r$  from zero, the analysis of variance technique provides an exact test which is applicable when the sample is either small or large.

Let us suppose that  $Y$  is a dependent variable which is to be predicted by a multiple regression equation containing  $m$  independent variables designated by  $X$ 's. The prediction equation may be written as

$$Y' = A + B_1X_1 + B_2X_2 + \cdots + B_mX_m$$

in which the  $B$ 's are the regression coefficients. The deviation of any individual's  $Y$  score from the mean  $Y$  can be expressed as the sum of 2 parts: the deviation of his  $Y$  from his predicted value plus the deviation of the predicted value from the mean of the  $Y$ 's, thus,

$$(Y - \bar{Y}) = (Y - Y') + (Y' - \bar{Y})$$

If we square both sides and sum over all cases, we have

$$\Sigma(Y - \bar{Y})^2 = \Sigma(Y - Y')^2 + \Sigma(Y' - \bar{Y})^2$$

which is exactly analogous to the breakdown used in connection with the test of the linear correlation coefficient. One part has to do with residuals about the regression *plane*, the other with variations in the predicted values. The cross-product term again vanishes—it can be shown that there is no correlation between residuals and predicted values.

As previously, we label the  $\Sigma(Y - Y')^2$  as the residual sum of squares and  $\Sigma(Y' - \bar{Y})^2$  as the regression sum of squares. The total sum of squares will, of course, have  $N - 1$  degrees of freedom. The residual sum of squares will lose  $df$ 's according to the number of constants in the regression equation. We have the constant  $A$ , and the number of  $B$  constants is  $m$ ; hence  $df = N - (m + 1) = N - m - 1$  for the residual term. The reader who does not immediately see the reasonableness of this should consider the case of 1 dependent and 2 independent variables with varying scores on  $N = 3$  cases. Imagine that the 3 scores

for each case can be used to locate a point for each in three-dimensional space, and then think of fitting an ordinary plane to these 3 points. Obviously, the plane can be made to pass through all 3; hence the prediction would be perfect, and there would be no freedom for any of the 3 points to vary from the plane. That is, with  $N = 3$  (and with variation on all 3 variables), the multiple derived therefrom must be unity.

Now, as to the  $df$  for the regression or prediction sum of squares, we note that for a fixed set of values for the  $X$ 's the variation of this term must depend upon the slopes of the regression plane or upon the  $B$ 's. There being  $m$   $B$ 's, there are  $m$  ways in which this sum can vary; therefore  $df = m$ . This is, it will be noted, an extension of the argument used to explain why  $df = 1$  for testing the linear correlation coefficient. If our  $df$  determinations are correct, we should have  $(N - m - 1) + m$  adding to  $N - 1$ , which is seen to be the case.

In Chapter 11 it was pointed out that the multiple correlation coefficient can be defined as

$$r^2_{1.23 \dots} = 1 - \frac{\sigma^2_{1.23 \dots}}{\sigma^2_1}$$

in which  $\sigma^2_{1.23 \dots}$  represents the residual variance and  $\sigma^2_1$  is the variance for the dependent variable. Since the residual variance plus the predicted variance adds to the total, the multiple  $r$  can also be expressed as the ratio of the predicted to the total variance. (Note that we are here speaking of variances, not estimates.) By definition, the residual variance is  $\Sigma(Y - Y')^2/N$ , the predicted variance is  $\Sigma(Y' - \bar{Y})^2/N$ , and the total variance is  $\Sigma\Sigma(Y - \bar{Y})^2/N$ . We may therefore write the multiple correlation coefficient, using  $R$  in order to avoid subscripts, as

$$R^2 = 1 - \frac{\Sigma(Y - Y')^2/N}{\Sigma\Sigma(Y - \bar{Y})^2/N}$$

from which it is readily seen that

$$\Sigma(Y - Y')^2 = (1 - R^2)\Sigma\Sigma(Y - \bar{Y})^2$$

From the alternate way of regarding multiple correlation, we have

$$R^2 = \frac{\Sigma(Y' - \bar{Y})^2/N}{\Sigma\Sigma(Y - \bar{Y})^2/N}$$

which leads to  $\Sigma(Y' - \bar{Y})^2 = R^2\Sigma\Sigma(Y - \bar{Y})^2$ .

Thus the sums of squares have their equivalents in terms of  $R$ , and consequently they may be computed by way of  $R$ . The computation of these sums directly would be a hammer-and-tongs approach which would involve the laborious task of predicting by means of the regression equation the  $Y$  for each individual.

The foregoing may be assembled in a schematic variance table, like Table 41. As in testing the significance of the ordinary corre-

Table 41. VARIANCE SETUP FOR TESTING SIGNIFICANCE OF MULTIPLE CORRELATION COEFFICIENT

Source	Sum of Squares	Equivalent	df	Estimate
Regression	$\Sigma(Y' - \bar{Y})^2 = R^2 \Sigma(Y - \bar{Y})^2$		$m$	$s_p^2$
Residual	$\Sigma(Y - Y')^2 = (1 - R^2) \Sigma(Y - \bar{Y})^2$		$N - m - 1$	$s_r^2$
Total	$\Sigma(Y - \bar{Y})^2$		$N - 1$	

lation coefficient, we set the null hypothesis to the effect that the estimate based on the regression sum of squares will differ from that based on the residual sum only because of chance sampling errors. The null hypothesis implies that, if the entire population were measured, the correlation of the dependent variable with each independent variable would be zero. Now, when a sample is drawn from such a population, the  $r$ 's will vary more or less from zero with the result that the multiple  $R$  will likewise differ from zero. If the conditions of the null hypothesis hold true, the sampling distribution of  $s_p^2/s_r^2$  follows that of the  $F$  distribution with appropriate degrees of freedom. Note that

$$\begin{aligned}
 F &= \frac{s_p^2}{s_r^2} = \frac{\Sigma(Y' - \bar{Y})^2/m}{\Sigma(Y - Y')^2/(N - m - 1)} \\
 &= \frac{R^2 \Sigma(Y - \bar{Y})^2/m}{(1 - R^2) \Sigma(Y - \bar{Y})^2/(N - m - 1)} \\
 &= \frac{R^2/m}{(1 - R^2)/(N - m - 1)}
 \end{aligned}$$

hence  $F$  is a ratio which depends upon  $R$  and the  $df$ 's. If the numerator is less than the denominator, we may conclude without reference to the table of  $F$  that  $R$  is insignificant. When the numerator is the larger, one judges the significance of  $F$  by entering the table of  $F$  with  $n_1 = m$  and  $n_2 = N - m - 1$ . Once  $R$  has been computed, the calculations involved in checking its significance are so simple that an example would be humdrum.

In the chapter on multiple correlation, it was pointed out that  $R$  as computed tends to have a positive bias, the extent of which could be judged by formula (75). This formula can readily be derived by the use of estimated residual and trait variances in place of actual variances in formula (70). Best or unbiased estimates lead to an unbiased  $R$ , or provide an unbiased estimate of the population value of  $R$ . Formula (75) gives this improved estimate, but the improvement is negligible except when  $N$  is small, or when  $m$  is large relative to  $N$ . It should be stressed that neither the analysis of variance check on the significance of  $R$  nor the improved estimate of  $R$  allows for the fallacy involved in multiple correlation work when from among a large number of variables a few are chosen for inclusion in the analysis because they show correlation with the criterion. Such selection tends to capitalize on  $r$ 's which are among the highest partly because of chance errors.

A practical question of considerable importance arises when one wonders whether the inclusion of additional variables in the multiple regression equation leads to a significant increase in the accuracy of prediction or when one wishes to know whether the dropping of certain variables results in a significant decrease in the amount of variance predicted. The inclusion of additional variables in the equation always tends to reduce the error of estimate somewhat and leads to an increase in  $R$ . Can it be said that the increase in  $R$  possesses statistical significance?

Let  $R_1$  be the multiple based on  $m_1$  independent variables and  $R_2$  be the value based on  $m_2$  variables *selected from among* the  $m_1$  variables. To test the significance of the difference between  $R_1$  and  $R_2$ , we take

$$F = \frac{(R_1^2 - R_2^2)/(m_1 - m_2)}{(1 - R_1^2)/(N - m_1 - 1)}$$

with  $n_1 = m_1 - m_2$  and  $n_2 = N - m_1 - 1$ . If  $F$  falls beyond

the .01 point, we can safely assume that the apparent gain in using the additional variable or variables possesses statistical significance.

### INTRACLASS CORRELATION

Suppose we wish to specify the degree of resemblance of twins in terms of a correlation coefficient. We have measurements on just 1 variable, and if we attempt to make a scatter diagram we are faced with the problem of deciding which member of a pair,  $A$  or  $A'$ , to assign to one axis and which to the other. This can be resolved by a double entry scheme: each pair is entered twice,  $A$  as  $X$  and  $A'$  as  $Y$ , and then  $A'$  as  $X$  and  $A$  as  $Y$ . An  $r$  calculated from the double entry (symmetrical) table suffers from a slight bias, which may be avoided by using the formula given below.

In general, if we have  $k$  families (or groups or classes) with  $m$  cases per family, the degree of resemblance can be specified by the intraclass correlation coefficient, computable by

$$r' = \frac{s^2_b - s^2_w}{s^2_b + (m - 1)s^2_w}$$

in which we have variance estimates for between families (groups or classes) and for within families. If  $F = s^2_b/s^2_w$  is significant we have evidence for a significant positive  $r'$ . Note that if there is no within-family variation,  $r'$  becomes unity. Note also that  $r'$  may be negative, but since in practice  $s^2_w$  will rarely be significantly larger than  $s^2_b$ , one is seldom confronted with the necessity for trying to interpret a negative intraclass correlation.

When the number of cases per family varies, the average of the  $m_r$  values is used in place of  $m$  in the foregoing formula for  $r'$ . This does not affect the  $F$  test as a way of judging the significance of the correlation.

The distinguishing characteristic of an intraclass correlation situation is that we have  $k$  sets of scores on just 1 variable with no way of ordering the scores within a set (a sort of interchangeability). It is obvious that  $r'$  can be used to describe group resemblance, regardless of how the groups have been defined.

## Analysis of Variance: Complex

In the previous chapter an explanation of the fundamental idea of the analysis of variance technique was attempted, and applications to relatively simple situations were given. In general, these situations involved the testing of the significance of the over-all variation of the means for several groups, the groups differing on the basis of a single classificatory principle. Such setups are sometimes referred to as *single variable experiments*, by which is meant that groups differing in *one* known respect are compared on a dependent variable. For example, income might be considered a variable which is dependent in part on amount of education, which accordingly becomes the independent, single variable for classifying individuals into groups. Or it might be that the classificatory variable is subject to experimental manipulation, and we wish to determine whether variations thereof will lead to performance or response differences. The Wright experiment cited in Chapter 15 is an example of this.

There are times when it is not only feasible but advisable to design the experimental setup so as to make one set of data serve for the testing of hypotheses regarding the separate influence of two or more independent variables. This type of thing has been done for a long time in psychological research wherein it has been possible to classify a total group first one way, then another, and perhaps a third way. For example, in order to determine some of the possible correlates of measured intelligence, we may classify a group of children into urban, suburban, and rural groups; then, ignoring this basis for grouping, we may classify them as to occupational level of father; or the classification may be by sex or by grade location or by age. Such a procedure in which one variable is considered at a time is tantamount to the single variable setup,



even though the same batch of data is made to answer questions about the effects of different independent variables.

Now it is obvious that, in studying factors associated with intelligence, we could make a double classification by classifying our cases simultaneously on two of the variables, or a triple classification by using three variables, etc. Consider for the moment a double classification based on the three rural-urban categories and on sex. This would lead to the assigning of the cases to six groups, each of which would have a mean IQ. Instead of having three means for groupings on the basis of the rural-urban characteristic, we would now have two sets of such means, one set for each sex. Instead of two means for the total group classified by sex, we would have three sets of sex means, a set for each of the three residence categories.

This type of breakdown and similar ones where percentages instead of means are involved were utilized in psychological research long before the advent of the analysis of variance technique. The further breakdown of each sex group for residence status (or of residence groups for sex) is made in order to see whether rural-urban differences hold for the sexes separately (or whether the sex differences are similar for each of the separate residence groups). Although researchers were not confined to the single variable approach before the invention of the variance technique, they were definitely limited in the possible statistical treatment of their data. Now that we have the analysis of variance method, we have an adequate statistical technique for checking such hypotheses as can be formulated concerning the influence of not only one but two or more variables. The advantages of using analysis of variance for such situations may be briefly mentioned.

First, as we have already seen, it provides an over-all test of the significance of the difference between two or more means when either large or small samples are involved.

Second, we shall soon see that it leads to a definitely improved estimate of sampling error when double or triple or higher-order classification is involved. For instance, when the older method is used to check the significance of the difference between the two sex means for the total group, the determination of the sampling error makes no allowance for likely heterogeneity in intelligence associated with residence status. The variance method permits a

refined estimate of error by allowing for variation due to one or more variables when one is testing the differences between groups classified on the basis of some other variable.

Third, the variance technique provides a means of testing whether the influence of one independent variable on the dependent variable is similar for subgroups formed on the basis of a second independent variable. In a sex-by-residence analysis of IQ's, the breakdown of each residence group by sex will likely show that the sex differences are not exactly the same for the three groups and that rural-suburban-urban differences are not exactly alike for the separate sex groups. Such inconsistencies as seem apparent from examination of the six cell means may not be real for the simple reason that random sampling errors are present. Before the development of the variance technique there was no way of testing such apparent inconsistencies, except when each classificatory characteristic led to just two categories.

This last point has to do with what has been termed *interaction*, a concept which is not easily understood. Rather than provide a detailed discussion now of what is meant by interaction, we will give a simple illustration. Suppose it has been found that one learning method has a distinct advantage over a second method, but that, when the data are broken down for two recall intervals, the superiority of the first method seems to hold only for those with the shorter recall interval. This failure of the first method to be consistently better becomes an example of interaction. Before concluding that there is evidence for real interaction, one needs to apply a statistical test. For such a simple breakdown, one could compute the difference between the first and second method means, and the standard error of the difference, for those with the short recall interval; likewise, for those with the long interval; then one could determine the difference between the differences and its standard error and therefrom obtain either a critical ratio or a  $t$  as a test of inconsistency. But, when one thinks of a situation with three methods and three or four recall intervals, it is immediately obvious that such a simple test cannot be applied.

It is the purpose of this chapter to present the methods of analysis to be used when classification into groups is made on the basis of two or more variables. These extensions, which are restricted by the underlying assumptions of normality and homo-

generality of certain variances, are applicable for either large or small samples and are particularly helpful with small samples when it seems imperative that we "get the most out of the available data."

### DOUBLE OR 2-WAY CLASSIFICATION

Suppose that the individuals (or their scores) are classifiable into  $C$  groups on the basis of one characteristic or variable and into  $R$  groups on the basis of a second variable. This would lead to a table with  $RC$  cells. Let us first examine the setup where we have only  $RC$  scores, i.e., one score for each cell. It is convenient to let  $X_{rc}$  stand for the score in the  $r$ th row and  $c$ th column of such a table. The score in the first row (from the top) and third column would be symbolized as  $X_{13}$ . The general pattern of labeling the scores is set forth in Table 42, which also includes along the margins a symbol for the several possible row and column means. Note that the first subscript identifies the row and the

Table 42. SCHEMA FOR LABELING SCORES AND MEANS FOR GROUPS, DOUBLE CLASSIFICATION

	1	2	3	$c$	$C$	
1	$X_{11}$	$X_{12}$	$X_{13}$	$X_{1c}$	$X_{1C}$	$\bar{X}_{1.}$
2	$X_{21}$	$X_{22}$	$X_{23}$	$X_{2c}$	$X_{2C}$	$\bar{X}_{2.}$
3	$X_{31}$	$X_{32}$	$X_{33}$	$X_{3c}$	$X_{3C}$	$\bar{X}_{3.}$
$r$	$X_{r1}$	$X_{r2}$	$X_{r3}$	$X_{rc}$	$X_{rC}$	$\bar{X}_{r.}$
$R$	$X_{R1}$	$X_{R2}$	$X_{R3}$	$X_{Rc}$	$X_{RC}$	$\bar{X}_{R.}$
	$\bar{X}_{.1}$	$\bar{X}_{.2}$	$\bar{X}_{.3}$	$\bar{X}_{.c}$	$\bar{X}_{.C}$	$\bar{X}$

second the column to which a score belongs. The scheme used in denoting means should be grasped. Thus  $\bar{X}_{.2}$  is the mean for the second column, whereas  $\bar{X}_{2.}$  is the mean for the second row. The "dot" in the subscript indicates the direction of the summing for computing a mean—to get  $\bar{X}_{.2}$  we sum  $X_{r2}$  scores with  $r$  taking on values running from 1 to  $R$ .

The deviation of any score,  $X_{rc}$ , from the total mean can be expressed in terms of the deviation of its row mean from the total mean,  $(\bar{X}_r - \bar{X})$ , plus the deviation of its column mean from the total mean,  $(\bar{X}_{\cdot c} - \bar{X})$ , plus a sort of remainder term which represents an individual variation over and above that due to the groups to which the score belongs. To secure an expression for this term, we note that by definition the term must be the part of the score deviation (from the total mean) left over after the sum of the two parts specified above have been subtracted. Accordingly, we have

$$(X_{rc} - \bar{X}) - [(\bar{X}_r - \bar{X}) + (\bar{X}_{\cdot c} - \bar{X})]$$

which simplifies to

$$(X_{rc} - \bar{X}_r - \bar{X}_{\cdot c} + \bar{X})$$

We may therefore write the following identity:

$$(X_{rc} - \bar{X}) = (\bar{X}_r - \bar{X}) + (\bar{X}_{\cdot c} - \bar{X}) + (X_{rc} - \bar{X}_r - \bar{X}_{\cdot c} + \bar{X})$$

With  $r$  running from 1 to  $R$ , and  $c$  taking values from 1 to  $C$ , there will, of course, be  $RC$  individual deviations. We need the sum of their squares, which sum will involve the squares of the three parts, plus three cross-product terms that can be shown to vanish when summed. It may be instructive to indicate how the sum of squares for all  $RC$  scores can be set up. Suppose we begin by writing the squares of the deviations for scores in the first column. Each of these squares will involve cross-product terms, which we shall here ignore except for a plus sign to indicate their existence. We have for the first-column scores:

$$(X_{11} - \bar{X})^2 = (\bar{X}_1 - \bar{X})^2 + (\bar{X}_{\cdot 1} - \bar{X})^2 + (X_{11} - \bar{X}_1 - \bar{X}_{\cdot 1} + \bar{X})^2 + \dots$$

$$(X_{21} - \bar{X})^2 = (\bar{X}_2 - \bar{X})^2 + (\bar{X}_{\cdot 1} - \bar{X})^2 + (X_{21} - \bar{X}_2 - \bar{X}_{\cdot 1} + \bar{X})^2 + \dots$$

$$(X_{r1} - \bar{X})^2 = (\bar{X}_r - \bar{X})^2 + (\bar{X}_{\cdot 1} - \bar{X})^2 + (X_{r1} - \bar{X}_r - \bar{X}_{\cdot 1} + \bar{X})^2 + \dots$$

$$(X_{R1} - \bar{X})^2 = (\bar{X}_R - \bar{X})^2 + (\bar{X}_{\cdot 1} - \bar{X})^2 + (X_{R1} - \bar{X}_R - \bar{X}_{\cdot 1} + \bar{X})^2 + \dots$$

The summing of these squares of deviations for scores of column 1 involves  $R$  cases, i.e.,  $r$  runs from 1 to  $R$ ; hence we need a symbol

which denotes this fact. Let us use  $\sum_r$  for this purpose. Note that the second term on the right is constant for all  $R$  scores, which permits us to replace the summation sign by  $R$ .

The sum of the first column squares, and by analogy the sums for the other columns, can be written as:

1st col.:

$$\sum^r (X_{r1} - \bar{X})^2 = \sum^r (\bar{X}_{r.} - \bar{X})^2 + R(\bar{X}_{.1} - \bar{X})^2 + \sum^r (X_{r1} - \bar{X}_{r.} - \bar{X}_{.1} + \bar{X})^2$$

2nd col.:

$$\sum^r (X_{r2} - \bar{X})^2 = \sum^r (\bar{X}_{r.} - \bar{X})^2 + R(\bar{X}_{.2} - \bar{X})^2 + \sum^r (X_{r2} - \bar{X}_{r.} - \bar{X}_{.2} + \bar{X})^2$$

cth col.:

$$\sum^r (X_{rc} - \bar{X})^2 = \sum^r (\bar{X}_{r.} - \bar{X})^2 + R(\bar{X}_{.c} - \bar{X})^2 + \sum^r (X_{rc} - \bar{X}_{r.} - \bar{X}_{.c} + \bar{X})^2$$

Cth col.:

$$\sum^r (X_{rC} - \bar{X})^2 = \sum^r (\bar{X}_{r.} - \bar{X})^2 + R(\bar{X}_{.C} - \bar{X})^2 + \sum^r (X_{rC} - \bar{X}_{r.} - \bar{X}_{.C} + \bar{X})^2$$

We may now sum over the  $C$  columns, and for the results we will need double summation signs. Since the first right-hand term does not vary from column to column, its sum is merely  $C$  times its value. The second right-hand set of terms involves a constant times a variable; hence the constant  $R$  comes from under the summation sign. Finally we have the following expression for the sum of squares for the  $RC$  scores:

$$\begin{aligned} \sum^r \sum^c (X_{rc} - \bar{X})^2 &= C \sum^r (\bar{X}_{r.} - \bar{X})^2 + R \sum^c (\bar{X}_{.c} - \bar{X})^2 \\ &\quad + \sum^r \sum^c (X_{rc} - \bar{X}_{r.} - \bar{X}_{.c} + \bar{X})^2 \quad (101) \end{aligned}$$

The reader who is worried about whether the cross-product terms really vanish should note that for the  $c$ th column the product term

$$\sum^r (\bar{X}_{r.} - \bar{X})(\bar{X}_{.c} - \bar{X}) = (\bar{X}_{.c} - \bar{X}) \sum^r (\bar{X}_{r.} - \bar{X})$$

vanishes because  $\sum^r (\bar{X}_{r.} - \bar{X}) = 0$ . The other two cross-product sums have as one factor the remainder or residual term; we have already had examples of a general principle that product terms involving residuals vanish.

From formula (101) we see that the total sum of squares can be broken into three additive components: between row means with  $R - 1$  degrees of freedom, between column means with  $df$  of  $C - 1$ , and a remainder. The degrees of freedom for the last part can be ascertained by a principle analogous to that used for



getting the  $\chi^2$   $df$  for contingency tables. The marginal means constitute restrictions on the deviation score entries in the rows and columns when deviation scores for  $(R - 1)(C - 1)$  cells are filled in, the rest of the entries become fixed; hence  $df = (R - 1)(C - 1)$ . Note that the  $df$ 's for the 3 parts sum to the  $df$  for the total sum of squares or  $RC - 1$ .

Dividing the 3 sums of squares by their  $df$ 's leads to 3 variance estimates,  $s^2_r$  for that based on rows,  $s^2_c$  for columns, and  $s^2_e$  for that based on the remainder, sometimes called error, sum of squares. We have 2 null hypotheses: that the row means are chance variations from 1 population mean, and that the column means are also variations from 1 population mean. As in the simpler situation, if the estimate based on rows is larger than expected on the basis of chance, it follows that there are real differences between the population means for the groups defined by the rows; likewise, for column means.

In testing the significance of either of the 2 between-groups variances when the  $RC$  scores belong to  $RC$  individuals, we use the remainder variance estimate as the denominator of the  $F$  ratio. This involves an assumption, to be discussed below under the heading "Choice of error term," p. 303. For testing the variation of row means, we have  $F = s^2_r/s^2_e$  with  $n_1 = R - 1$  and  $n_2 = (R - 1)(C - 1)$ . For column means,  $F = s^2_c/s^2_e$  with  $n_1 = C - 1$  and  $n_2 = (R - 1)(C - 1)$ . If an  $F$  so defined happens to be less than unity, we know at once without reference to the table for  $F$  that the variations of the given means are insignificant. Note that, since the error variance used in the denominator is a residual after the parts of the total associated with between-row and between-column variations have been subtracted, it follows that we are using as our error term a variance which has been freed of the influence of heterogeneity with respect to the 2 classificatory variables being investigated.

For many situations involving double classification, it would seem that the method just outlined would be definitely limited in usefulness because no provision has been made for increasing the size of the sample except by using finer grouping on one or both of the independent variables. Finer grouping would be possible, though not always feasible or desirable, for some classificatory variables, such as degree of illumination or amount of



education or size of type, but for other bases for forming groups there are definite limits on the number of groups. For example, in the study of reaction time the number of possible groupings for sense modality is limited. Actually, the number of cases can be increased by having additional individuals assigned to each of the  $RC$  cells. Before taking up this needed modification of the setup, we shall discuss certain specific situations where the scheme as presented is of practical use. We are not ignoring the possibility that sometimes  $RC$  cases are enough for testing hypotheses even when both  $R$  and  $C$  are as small as 4 or 5.

### SIGNIFICANCE OF THE DIFFERENCES BETWEEN CORRELATED MEANS

Suppose that the  $RC$  scores are for  $R$  individuals working under  $C$  different conditions. The mean of a row would be for an individual, and the mean of a column would be for a specified condition. Let us consider the limiting case of  $C = 2$ . The between-columns sum of squares,  $R\sum^c(\bar{X}_{\cdot c} - \bar{X})^2$ , may be written as

$$R(\bar{X}_{\cdot 1} - \bar{X})^2 + R(\bar{X}_{\cdot 2} - \bar{X})^2$$

which we have already shown (p. 260) reduces to  $(R/2)(\bar{X}_{\cdot 1} - \bar{X}_{\cdot 2})^2$ , or to a function of the difference between the two means.

Let us next examine the remainder or error term. If we turn back to p. 286, where we summed over columns, we readily see that the remainder sum can be expressed as

$$\sum^r(X_{r1} - \bar{X}_r - \bar{X}_{\cdot 1} + \bar{X})^2 + \sum^r(X_{r2} - \bar{X}_r - \bar{X}_{\cdot 2} + \bar{X})^2$$

in which the  $c$  of formula (101) has the explicit values of 1 and 2. Now the mean of any row, say the  $r$ th, is merely the mean of  $C = 2$  scores; i.e.,  $\bar{X}_r = (X_{r1} + X_{r2})/2$ , and the total mean must be the average of the two column means, or  $\bar{X} = (\bar{X}_{\cdot 1} + \bar{X}_{\cdot 2})/2$ . Making these substitutions, we have

$$\begin{aligned} \sum^r \left( X_{r1} - \frac{X_{r1} + X_{r2}}{2} - \bar{X}_{\cdot 1} + \frac{\bar{X}_{\cdot 1} + \bar{X}_{\cdot 2}}{2} \right)^2 \\ + \sum^r \left( X_{r2} - \frac{X_{r1} + X_{r2}}{2} - \bar{X}_{\cdot 2} + \frac{\bar{X}_{\cdot 1} + \bar{X}_{\cdot 2}}{2} \right)^2 \end{aligned}$$

which simplifies to

$$\frac{1}{4} \sum^r (X_{r1} - X_{r2} - \bar{X}_{\cdot 1} + \bar{X}_{\cdot 2})^2 + \frac{1}{4} \sum^r (X_{r2} - X_{r1} - \bar{X}_{\cdot 2} + \bar{X}_{\cdot 1})^2$$

These two terms become identical when we change the signs within the second parentheses, which change is permissible since the square of a function is the same as the square of its negative, e.g.,  $(a)^2 = (-a)^2$ . Hence we have

$$\frac{1}{2} \sum^r [(X_{r1} - X_{r2}) - (\bar{X}_{\cdot 1} - \bar{X}_{\cdot 2})]^2$$

Now the first parentheses term is the difference between any individual's two scores, say  $D_r$ , and the second is the difference between the two column means, which difference it will be recalled is the same as the mean of the differences,  $\bar{D}$ . We have finally the remainder sum of squares as  $\frac{1}{2} \sum^r (D_r - \bar{D})^2$ , or one-half the sum of the squares of the difference scores about the mean difference.

The  $F$  for comparing two column means becomes

$$F = \frac{\frac{R}{2} (\bar{X}_{\cdot 1} - \bar{X}_{\cdot 2})^2}{\frac{s_c^2}{s_d^2} = \frac{1}{\frac{\frac{1}{2} \sum^r (D_r - \bar{D})^2}{R - 1}}}$$

with  $n_1 = 1$  and  $n_2 = R - 1$ . This reduces to

$$F = \frac{(\bar{X}_{\cdot 1} - \bar{X}_{\cdot 2})^2}{\frac{\sum^r (D_r - \bar{D})^2}{R(R - 1)}}$$

which the reader will recognize as  $t^2$  for comparing the difference between means based on sets of correlated scores with the standard error of the mean difference estimated by formula (28), p. 108.

We have seen in Chapter 6 that in testing the difference between the means of correlated scores we can, for the large sample situation, determine the needed sampling error either from the distribution of differences between paired scores or by means of the standard error of the difference formula with the correlational

term included. The important thing to note is that the analysis of variance technique provides a method for testing the significance of the difference between two or *more* means based on sets of correlated scores. The scores may be correlated either because they are based on the *same* individuals working under *C* conditions or having *C* trials on some stunt, or because siblings or litter mates are involved (each of the *C* groups containing one case from each of *R* families), or because we started with *R* sets of matched individuals, one from each set being assigned to the several *C* groups. After and only after it has been found that the *F* for the *C* column means is significant are we justified in using the critical ratio or *t* technique to test the significance of the difference between any two of the *C* means.

The *F* just discussed has to do with column means. What of the row means for the given setup? The means of the *R* rows represent the mean performance of each of the several individuals, and a test of the significance of the estimate of variance based on the between-row sum of squares becomes a test of the significance of individual differences. Since it is known that individuals do differ on practically all psychological variables, such a test is usually a trivial test of the obvious, and hence it is seldom needed. We may, however, have the situation in which we wonder whether individual variation is significant in the light of known measurement or response errors. To this question we now turn.

### RELIABILITY OF MEASUREMENT

Suppose the scores in each row represent either the performance of an individual on different forms of a scale or *C* measurements for a given variable. The column means would be the means for the forms or successive sets of measurements, and the test of the significance between column means would be a test of the difference between the several form means or of the difference between the means for the *C* successive sets of trials. For form means or for trial means,  $F = s^2_c/s^2_e$ , as outlined above, provides an over-all test of the significance of these *correlated* means.

In order better to understand the meaning, in this situation, of an *F* based on  $s^2_r$ , let us again take the limiting case of *C* = 2; e.g., suppose two forms of a test have been administered to *R* individuals. The algebra is simplified and an interesting, clear-

cut result emerges if we assume that the two forms yield exactly the same means, i.e., that  $\bar{X}_{.1} = \bar{X}_{.2} = \bar{X}$ . Then the remainder sum of squares,

$$\sum^r \sum^c (X_{rc} - \bar{X}_{r.} - \bar{X}_{.c} + \bar{X})^2$$

becomes

$$\sum^r \sum^c (X_{rc} - \bar{X}_{r.})^2$$

This can be written without the double summation sign as

$$\sum^r (X_{r1} - \bar{X}_{r.})^2 + \sum^r (X_{r2} - \bar{X}_{r.})^2$$

Since the mean of each row is simply the average of two scores, i.e.,

$$\bar{X}_{r.} = \frac{X_{r1} + X_{r2}}{2}$$

the above can be written as

$$\sum^r \left( X_{r1} - \frac{X_{r1} + X_{r2}}{2} \right)^2 + \sum^r \left( X_{r2} - \frac{X_{r1} + X_{r2}}{2} \right)^2$$

which by a little algebraic manipulation reduces to

$$\frac{1}{2} \sum^r (X_{r1} - X_{r2})^2$$

Since we have assumed that the form means are equal, the difference scores in this expression will have a mean of zero. Therefore, if we divide the sum of the squared differences by  $R$ , the number of individuals, we will have the variance of the distribution of differences, which we symbolize by  $\sigma_{DD}^2$ .

It follows that

$$\frac{1}{2} \sum^r (X_{r1} - X_{r2})^2 = \frac{R}{2} \sigma_{DD}^2$$

Now it can be shown by easy algebra (see pp. 83-84) that

$$\sigma_{DD}^2 = \sigma_1^2 + \sigma_2^2 - 2r_{12}\sigma_1\sigma_2$$

in which the  $\sigma$ 's are measures of variation for forms 1 and 2 respectively, and  $r_{12}$  is the correlation between forms. If we make the usual assumption that the two forms are so nearly comparable that we can replace  $\sigma_1$  and  $\sigma_2$  by  $\sigma$ , we have

$$\sigma_{DD}^2 = \sigma^2 + \sigma^2 - 2r_{12}\sigma\sigma = 2\sigma^2(1 - r_{12})$$

Then  $(R/2)\sigma_{DD}^2$  becomes  $R\sigma^2(1 - r_{12})$ . But  $r_{12}$  defines, and is, the reliability coefficient, and hence  $\sigma^2(1 - r_{12})$  is the error of measurement variance,  $\sigma_e^2$ , so that we finally have the remainder sum of squares equal to  $R\sigma_e^2$ .

Thus, under our simplifying assumptions of equal form means and equal form variances, assumptions which are usually made in connection with test reliability, we see that the remainder term is directly associated with the familiar error of measurement variance. The remainder term as actually computed from the sum of squares includes an adjustment for possibly differing form means but no allowance for differing form variances, so it will not exactly equal  $R\sigma_e^2$ . The remainder sum of squares does, however, lead to an estimate of the error of measurement variance, not only in the situation where we have an analysis based on two forms but also where three or more forms are involved; accordingly, when we test the significance of the variance for between-row means, we are actually asking whether the individual differences are significant in light of the variability due to measurement errors.

Since the reliability coefficient is a function of the error variance relative to the observed trait variance, it follows that a significant between-individuals variance is evidence for statistically significant reliability. But one cannot conclude from this that the test or instrument possesses satisfactory reliability since coefficients as low as .20 or .30 or even .10 can be statistically different from zero if  $R$  is sufficiently large. The author does not recommend this approach to the question of the reliability of measurement for the simple reason that it is more important to know *how* reliable a test is or how near its reliability approaches unity than to know only that it *is* reliable in the sense of yielding a coefficient significantly different from zero.

This possible application of the variance technique, however, points up the fact that it is sometimes meaningful to speak of the remainder variance as "error" variance. In a wider sense, the remainder variance can be thought of as the uncontrolled variation which contributes to the variation of the means of the groups being compared. Now a little reflection leads one to the conclusion that the sources of error in research are many and varied. Sometimes instrumental and/or measurement errors loom large, some-



times the error associated with the sampling of individuals is paramount, at other times the intraindividual variation is sizable, and frequently if the sources of variation are unknown the term experimental error is used as a catchall. When a particular variance estimate is referred to as *the* error variance to be used as the denominator of the  $F$  ratio, the "error" may be any one of or a combination of the many types of error. In this sense, the variance estimate based on the remainder sum of squares may be the error variance even for those situations where we have classifications into  $R$  groups rather than as  $R$  individuals, but as will presently be seen the term which we are now calling the remainder may not always be the one to utilize as "error." The within-groups variance estimate of the last chapter was an "error" variance for testing the significance of the between-groups variation. In more complex setups in the analysis of variance, judgment is required in choosing the appropriate error term.

Parenthetically, it might be pointed out that the test reliability problem can be tackled by the within- and between-groups variance estimates. Each person for whom we have two or more, say  $C$ , measurements yields a set of  $C$  scores, and the variation within such a set is partly a function of measurement errors; hence the over-all within-groups (intraindividual) variance estimate becomes an error term by which one may test the significance of the between-groups (between-individual) variance. Note that this within-groups or intraindividual approach will lead to an estimate of the error of measurement variance *without* an adjustment for possible differences in form means, and that it does not permit a test of the significance of the difference between form means, which is possible when the double classification scheme is utilized. Either of the two methods for determining whether the reliability is sufficient to possess statistical significance is applicable for an over-all evaluation of  $C$  forms or  $C$  successive measurements or trials. With  $C$  forms and  $R$  individuals, it is of interest to make a comparative layout of the two approaches, that based on the double classification scheme of this chapter and that based on the single classification procedure of the last chapter. Table 43 contains the essentials.

Note that both  $F = s^2_r/s^2_e$  and  $F = s^2_b/s^2_w$  provide tests of the significance of reliability by way of the significance of indi-



Table 43. TWO APPROACHES TO TEST RELIABILITY PROBLEM

Via Double Classification		Via Single Classification	
Variance Estimate	<i>df</i>	Variance Estimate	<i>df</i>
$s^2_r$	$R - 1$	$s^2_b$	$R - 1$
$s^2_c$	$C - 1$		
$s^2_e$	$(R - 1)(C - 1)$	$s^2_w$	$R(C - 1)$

vidual differences. The *df* for the estimate  $s^2_e$  is  $C - 1$  smaller than that for  $s^2_w$ , a trivial difference in the practical situation where  $C$  is seldom more than 2 or 3, and  $R$  is usually 100 or more, rarely as small as 25 or 50. Both  $s^2_e$  and  $s^2_w$  constitute estimates of the error of measurement variance, but  $s_e$ , because of the adjustment for differing form means, will be smaller than  $s_w$ . Whether either of these estimates is useful as indicating precisely the measurement error for a particular form depends upon the extent to which the standard deviations for the several forms are similar.

### COMPUTATIONAL ILLUSTRATION

The required computations for testing variation between column means and between row means will now be set forth. It makes no difference in the computational procedure whether we have  $RC$  individuals classified into  $R$  groups one way and  $C$  groups another way or  $R$  individuals with  $C$  scores each or  $R$  sets of  $C$  individuals matched or  $RC$  scores for just 1 individual.

The computation of the required sums of squares involves an extension of formulas (97), as follows:

$$\sum \sum (X_{rc} - \bar{X})^2 = \frac{1}{RC} [RC \sum \sum X_{rc}^2 - (\sum \sum X_{rc})^2] \quad \text{for total} \quad (102a)$$

$$R \sum (\bar{X}_{.c} - \bar{X})^2 = \frac{1}{RC} [C \sum (\sum X_{rc})^2 - (\sum \sum X_{rc})^2] \\ \text{for columns} \quad (102b)$$

$$C \sum (\bar{X}_{r.} - \bar{X})^2 = \frac{1}{RC} [R \sum (\sum X_{rc})^2 - (\sum \sum X_{rc})^2] \quad \text{for rows} \quad (102c)$$

The sum of squares for the remainder can be obtained by subtracting the sums for between columns and for between rows from the total sum of squares. Formulas (102) may look forbidding at first, but actually the sums based on raw scores are easily secured by following a plan on the work sheet. Sum each row, and write the sums on the right-hand margin; sum each column, and write the sums along the bottom margin. Summing down the right-hand margin gives the total sum, and summing across the bottom margin should give the same total sum. Square all scores and sum to get the first sum in (102a); square all the right-hand margin sums and then sum to get the first part of (102c); square all the bottom margin sums and then sum to get the first part of (102b).

The student may do well to sit down at a calculator and perform these operations with the scores in Table 44, which contains

Table 44. DATA FOR VISUAL ACUITY, 4 INDIVIDUALS, 3 DISTANCES  
(MONOCULAR, VERNIER METHOD, CODED SCORES) \*

Subjects	Distance (in Meters)			$\sum X_{rc}$	$\bar{X}_r$
	5	10	15		
1	13	29	17	59	19.7
2	4	9	19	32	10.7
3	8	30	37	75	25.0
4	9	27	53	89	29.7
$\sum X_{rc}$	34	95	126	255	
$\bar{X}_c$	8.5	23.7	31.5	21.2 = $\bar{X}$	

$$\sum \sum X_{rc} = 255$$

$$\sum \sum X_{rc}^2 = 7709$$

$$\sum (\sum X_{rc})^2 = 18,051$$

$$\sum (\sum X_{rc})^2 = 26,057$$

\* From Walker, E. L., *Factors in vernier acuity and distance discrimination*, Doctoral Dissertation, Stanford University, California, 1947.

visual acuity data on 4 (=  $R$ ) individuals for 3 (=  $C$ ) distances of the stimulus from the eye. Casual examination of the table indicates that acuity measures are influenced by distance. Do the means for the 3 distances differ significantly?

The required sums are also included in the table. Substituting these in the above formulas gives:

$\frac{1}{12}[12(7709) - (255)^2] = 2290.25$  for the total sum of squares

$\frac{1}{12}[3(26,057) - (255)^2] = 1095.50$  for between-columns sum of squares

$\frac{1}{12}[4(18,051) - (255)^2] = 598.25$  for the between-rows sum of squares

Subtracting the sum of the last 2 from the total gives 596.50 as the remainder sum of squares.

These results are assembled in Table 45 along with the *df*'s and the variance estimates. For the influence of distance we have

Table 45. VARIANCE TABLE FOR DATA OF TABLE 44

Source	Sum of Squares	<i>df</i>	Variance Estimate
Distance	1095.50	2	547.75
Subjects	598.25	3	199.42
Remainder	596.50	6	99.42
Total	2290.25	11	

$F = 547.75/99.92 = 5.51$ , which for  $n_1 = 2$  and  $n_2 = 6$  is significant at slightly better than the  $P = .05$  level (additional data in Walker's dissertation leave no doubt—distance does have an effect). This is a situation in which experimentally induced differences are so large that they can be demonstrated with only 4 cases.

#### DOUBLE CLASSIFICATION WITH MORE THAN ONE SCORE PER CELL

Suppose that we have  $m$  scores in each cell of schematic Table 42. This would lead to a mean for each cell, and about each such mean we would have the variation of  $m$  scores. The mean for the  $r$ th row would be the mean of all  $mC$  scores in the row, or the mean of the  $C$  cell means of the row; the mean of the  $c$ th column would be the mean of the  $mR$  scores in the column, or the mean of the cell means in the column; in the remainder term, previously defined as  $(X_{rc} - \bar{X}_r - \bar{X}_c + \bar{X})$ , we would replace  $X_{rc}$  by  $\bar{X}_{rc}$ . The total sum of squares for all  $mRC$  scores would include a between-column, a between-row, and a remainder component, plus an additional part which would involve the variation within

cells about the cell means. A convenient label for this new part would be  $\sum\sum(X_{rc} - \bar{X}_{rc})^2$ , in which it is understood that there are  $m$  such deviations in each cell. A more precise notation would be  $\sum\sum\sum(X_{irc} - \bar{X}_{rc})^2$ , in which  $X_{irc}$  is the  $i$ th score in the cell involving the  $r$ th row and  $c$ th column.

The variance table would take on the form indicated in Table 46, in which the term "remainder" has been replaced by "inter-

Table 46. VARIANCE SCHEMA FOR DOUBLE CLASSIFICATION WITH  $m$  SCORES PER CELL

Source	Sum of Squares	<i>df</i>	Variance Estimate
Rows	$m\sum^r(\bar{X}_{r.} - \bar{X})^2$	$R - 1$	$s^2_r$
Columns	$mR\sum^c(\bar{X}_{.c} - \bar{X})^2$	$C - 1$	$s^2_c$
Interaction	$m\sum^r\sum^c(\bar{X}_{rc} - \bar{X}_{r.} - \bar{X}_{.c} + \bar{X})^2$	$(R - 1)(C - 1)$	$s^2_{rc}$
Within cells	$\sum^r\sum^c\sum(X_{rc} - \bar{X}_{rc})^2$	$mRC - RC$	$s^2_w$
Total	$\sum^r\sum^c\sum(X_{rc} - \bar{X})^2$	$mRC - 1$	

action." Note that the first 2 sums of squares are simply  $m$  times the corresponding sums for 1 score per cell, and that the  $df$ 's for these sums and for the one corresponding to the remainder sum are not changed. The  $df$  for the within-cells sum depends upon the fact that there are  $m - 1$  degrees of freedom in each of the  $RC$  cells, which gives  $RC(m - 1) = mRC - RC$  as the  $df$ . We now have 4 estimates,  $s^2_r$ ,  $s^2_c$ ,  $s^2_{rc}$ ,  $s^2_w$ , of variance.

This simple modification of the setup for the analysis of variance leads to 2 definite advantages. We can increase the precision or dependability of our results by basing the analysis on more scores or cases, and we can test the possible significance of the interaction component. Before we discuss the first advantage, it is necessary that we consider the question of possible interaction, the exposition of which is facilitated by an example, which will also serve to illustrate the required computations.

The computational formulas are extensions of previously used formulas. A  $\Sigma X$  and  $\Sigma X^2$  is calculated for each cell. Summing the  $RC \Sigma X^2$  values gives  $\Sigma \Sigma X^2$  as the sum of all the  $mRC$  squared scores. Summing the  $\Sigma X$  values in each row gives  $\Sigma X_{rc}$ , and summing the  $\Sigma X$  values in each column gives  $\Sigma X_{rc}$ . These become sums along the margins, which marginal values sum down, and across, to the total sum of the  $mRC$  scores,  $\Sigma \Sigma X_{rc}$ . The sum of scores in any particular cell will be symbolized as  $\Sigma X_{rc}$ . The formulas are:

$$\text{Total sum of squares} = \frac{1}{mRC} [mRC \Sigma \Sigma X^2_{rc} - (\Sigma \Sigma X_{rc})^2] \quad (103a)$$

$$\text{Between-rows squares} = \frac{1}{mRC} [R \Sigma (\Sigma X_{rc})^2 - (\Sigma \Sigma X_{rc})^2] \quad (103b)$$

Between-columns squares

$$= \frac{1}{mRC} [C \Sigma (\Sigma X_{rc})^2 - (\Sigma \Sigma X_{rc})^2] \quad (103c)$$

$$\text{Within-cells squares} = \frac{1}{m} [m \Sigma \Sigma X^2_{rc} - \Sigma (\Sigma X_{rc})^2] \quad (103d)$$

The interaction sum of squares is obtained as the remainder when the numerical values of formulas (103bcd) are subtracted from the total sum of squares.

Table 47 contains data on learning with 2 variations as to practice sessions and 2 variations as to rest interval between trials. For each combination of conditions there are 20 ( $= m$ ) cases. The scores are recorded in a 2 by 2 or 4-cell table. Table 48 is a work-sheet layout in which are recorded sums of scores, sums of squared scores, and means, for cells and for the margins. The lower right corner contains values for the total group of 80 cases. For the sums of squares (of deviations) we have the following:

$$\text{Total: } \frac{1}{80} [80(7835) - (735)^2] = 1082.1875.$$

$$\text{Rows: } \frac{1}{80} [2(436^2 + 299^2) - (735)^2] = 234.6125.$$

$$\text{Columns: } \frac{1}{80} [2(341^2 + 394^2) - (735)^2] = 35.1125.$$

$$\text{Within cells: } \frac{1}{20} [20(7835) - (217^2 + 219^2 + 124^2 + 175^2)] = 782.4500.$$

$$\text{Interaction: } 1082.1875 - (234.6125 + 35.1125 + 782.4500) = 30.0125.$$

Table 47. CODED LEARNING SCORES (SUM OF SCORES ON 29TH AND 30TH TRIALS) FOR KOERTH PURSUIT ROTOR \*

Rest Interval	Practice Sessions							
	5(M T W Th F)				3(M W F)			
3 minutes	9	14	6	10	8	10	11	14
	10	15	10	11	9	7	9	10
	14	17	10	11	9	12	13	14
	10	7	8	15	12	13	7	17
	12	8	14	6	9	12	8	15
1 minute	2	6	1	9	11	12	9	7
	5	9	2	11	9	6	11	9
	14	1	1	8	6	8	11	12
	14	4	11	5	9	7	4	10
	6	8	2	5	13	6	7	8

\* Data from Renshaw, M. J., *The Effects of varied arrangements of practice and rest on proficiency in the acquisition of a motor skill*, Unpublished Doctor's Dissertation, Stanford University, California, 1947.

Table 48. SUMS AND MEANS FOR DATA OF TABLE 47

Rest Interval	Practice Session		Totals
	5(M T W Th F)	3(M W F)	
3 minutes	$\Sigma X_{11} = 217$	$\Sigma X_{12} = 219$	$\Sigma X_{10} = 436$
	$\Sigma X^2_{11} = 2543$	$\Sigma X^2_{12} = 2547$	$\Sigma X^2_{10} = 5090$
	$\bar{X}_{11} = 10.8500$	$\bar{X}_{12} = 10.9500$	$\bar{X}_{1.} = 10.9000$
1 minute	$\Sigma X_{21} = 124$	$\Sigma X_{22} = 175$	$\Sigma X_{20} = 299$
	$\Sigma X^2_{21} = 1102$	$\Sigma X^2_{22} = 1643$	$\Sigma X^2_{20} = 2745$
	$\bar{X}_{21} = 6.2000$	$\bar{X}_{22} = 8.7500$	$\bar{X}_{2.} = 7.4750$
Totals	$\Sigma X_{r1} = 341$	$\Sigma X_{r2} = 394$	$\Sigma \Sigma X_{ro} = 735$
	$\Sigma X^2_{r1} = 3645$	$\Sigma X^2_{r2} = 4190$	$\Sigma \Sigma X^2_{ro} = 7835$
	$\bar{X}_{.1} = 8.5250$	$\bar{X}_{.2} = 9.8500$	$\bar{X} = 9.1875$



The interaction sum of squares can also be calculated by direct substitution into the definition formula of Table 46, which will involve  $RC$  quantities to be squared, summed, and multiplied by  $m$ . We have

$$(10.85 - 10.90 - 8.525 + 9.1875)^2 = (.6125)^2$$

$$(10.95 - 10.90 - 9.85 + 9.1875)^2 = (-.6125)^2$$

$$(6.20 - 7.475 - 8.525 + 9.1875)^2 = (-.6125)^2$$

$$(8.75 - 7.475 - 9.85 + 9.1875)^2 = (.6125)^2$$

which when added and multiplied by 20 lead to 30.0125, or the value obtained by subtraction.

Any reader who is surprised that the above 4 values involved in computing the interaction sum of squares directly are numerically equal should ponder the fact that for the given situation the  $df$  for the interaction term is  $(2 - 1)(2 - 1)$  or 1.

Actually, the easiest way to compute the interaction sum of squares for a 2 by 2 table is to work with the 4 cell sums of scores. The formula is

$$\frac{1}{4m} (\Sigma X_{11} + \Sigma X_{22} - \Sigma X_{12} - \Sigma X_{21})^2$$

For this problem we have

$$\frac{1}{80} (217 + 175 - 219 - 124)^2 = \frac{1}{80} (49)^2 = 30.0125$$

The sums of squares and resulting variance estimates are brought together in Table 49. We have 4 variance estimates which

Table 49. ANALYSIS OF VARIANCE FOR PURSUIT LEARNING

Source	Sum of Squares	<i>df</i>	Variance Estimate
Rest interval (rows)	234.6125	1	234.6125
Sessions (columns)	35.1125	1	35.1125
Interaction	30.0125	1	30.0125
Individual differences (within cells)	782.4500	76	10.2954
Total	1082.1875	79	

for the given situation are all estimates of the same population variance under the null hypothesis conditions: no row effect, no column effect, and no interaction. It is appropriate for this table to use  $s^2_w$  as the denominator of  $F$  to test the row, the column, and the interaction effects. We have for interaction,  $F_{rc} = 30.0125/10.2954 = 2.92$ , which falls short of the  $F$  of about 4.0 required for significance at the .05 level. This indicates that the apparent failure of the 4 cell means to be consistent, in either direction, with the marginal means (or with each other) is attributable to chance fluctuations. For this particular problem the chance fluctuation is the sampling of individuals (plus a relatively small component having to do with errors of measurement).

Next consider the effect on pursuit learning of varying the rest interval and varying the sessions. For sessions we have  $F_c = 35.1125/10.2954 = 3.41$ , which is not large enough to lead us to reject the null hypothesis; but since nonrejection of the null hypothesis does not prove the hypothesis, we can conclude only that the effect, if it exists, is not large enough to be demonstrated by the number of cases used. The between-rows or rest-interval effect is highly significant as judged by  $F_r = 234.6125/10.2954 = 22.79$ , which is double the  $F$  needed for the .001 level of significance. Now the fact that the interaction is not significant permits us to conclude that the rest-interval effect is similar for 5 sessions and for 3 sessions per week. If the interaction had been significant, we would need to qualify our conclusion about the effect of the rest interval.

### ILLUSTRATIONS OF INTERACTION

Reference to actual examples of statistically significant interaction may help clarify its meaning. For this purpose we shall again use some data on visual acuity from the experiment by Walker.\* For visual acuity (low score, better acuity) by 2 methods of measurement (depth and vernier) with binocular and monocular vision, we have means as given in Table 50. The marginal means are markedly different, and it is readily seen that the cell means (each based on 108 determinations) are not consistent with the marginal values. The ratio of 1 to 3 for the binocular vs. monocular

\* Walker, E. L., *Factors in vernier acuity and distance discrimination*, Unpublished Doctor's Dissertation, Stanford University, California, 1947.

Table 50. VISUAL ACUITY: INTERACTION OF TYPE OF MEASUREMENT WITH EYES

	Depth	Vernier	Total
Binocular	.08	1.07	.57
Monocular	.24	1.50	.87
Total	.16	1.28	.72

means of .08 and .24 varies from the 2 to 3 ratio for the means on the right-hand margin, and the ratio of near 1 to 13 for the values of .08 and 1.07 differs from the 1 to 8 ratio of .16 to 1.28. In other words, the amount of difference between binocular and monocular acuity depends upon the type of measurement.

One variable investigated in the experiment was the distance of the stimulus from the subject. Since distance is an ordered

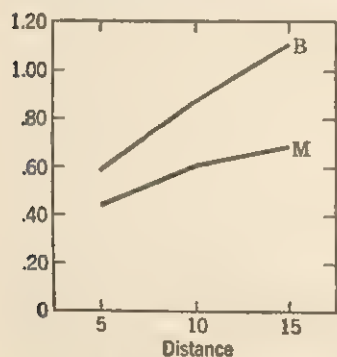


Fig. 15. Simple interaction: eyes by distance.

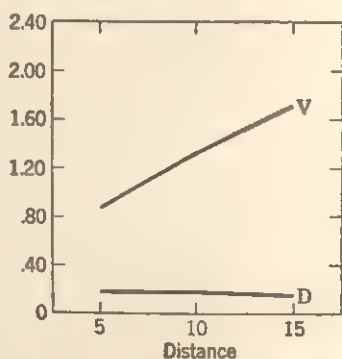


Fig. 16. Simple interaction: measures by distance.

variable, it is possible to picture the interaction by making a graph, with acuity as the ordinate and distance along the  $x$  axis. Figure 15 shows the relationship of acuity (average of the 2 types of measures) and the 3 distances used. Note the difference between the 2 curves—the significant interaction for eyes and distance actually means that the 2 curves are different. This lack of parallel behavior of curves is more striking in Fig. 16, which illustrates the interaction of measures with distance, for binocular and monocular combined. In this study there was also a significant variance for the subjects by distance interaction, from which

one concludes that the relationship between acuity and distance varies from person to person (see Fig. 17).

Walker also investigated the effect of stimulus rod width and size of aperture. A plot of the results for acuity (ordinate) against rod width (abscissa) for 3 apertures (*A* large, *B* medium, *C* small) is given in Fig. 18 as another possible example of interaction except that this time the apparent interaction is so slight as not to possess statistical significance. This being the case, it can be said that the effect of rod width is independent of aperture (and

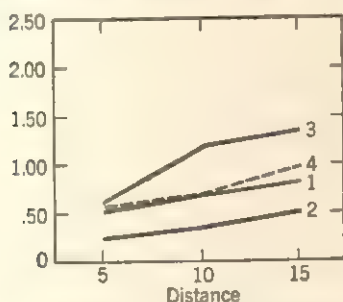


Fig. 17. Simple interaction: distance by subjects.

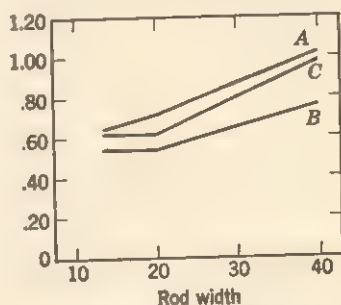


Fig. 18. Nonsignificant interaction: aperture by stimulus rod width.

vice versa). Contrast this with the possible conclusion, based on a highly significant *F*, that distance affects acuity. When we note the interaction effect depicted in Fig. 16, we see that such a conclusion does not hold at all for the depth measure. Thus, significant interaction always calls for a qualification, sometimes drastic, regarding a main effect. It is entirely possible for an effect to be in opposite directions for different conditions, and the over-all effect need not be significant for this to occur.

### CHOICE OF ERROR TERM IN 2-WAY CLASSIFICATION

Now that we have learned something about the meaning of interaction and have had a couple of examples which illustrate the computations and the way hypotheses can be tested, we must specifically consider an as yet unmentioned question: Which variance estimate is the correct one to use as the error term, that is, as the denominator for the *F* ratio? The answer depends upon

the mathematical model that is appropriate for a given situation. Three models have been set forth by the mathematical statisticians.† These are referred to as the components of variance model, the fixed constants model, and the mixed model. Let us define these for the 2-way classification setup.

We have the *components of variance model* when both classifications involve sampling. Such would be the case when rows stand for individuals and columns stand for judges (each of whom has rated each individual). The individuals and the judges are regarded as random samples from normally distributed populations: normal distribution of individuals with respect to the ratings and normal distribution for the rating characteristics of the judges.

We have a *fixed constants model* when no random sampling is involved so far as the bases of the classifications are concerned. Such is the case when the classifications depend upon such things as size, distance, time interval, degree of illumination, etc.; or on such unordered things as sense modality, sex, method, diagnostic group, etc. The setup in Table 47 involves the fixed constants model; neither the rest intervals nor the sessions were chosen at random.

We have a *mixed model* when one basis of classification involves sampling and the other fixed constants. Table 44 illustrates a typical mixed model, typical in that one basis of classification is individuals.

Each of the 3 models calls for precisely the same breakdown of the sum of squares and of the degrees of freedom, and each leads to 3 variance estimates plus a within-cells estimate in case we have more than 1 score per cell. It should be noted that the within-cells scores can stand for 2 kinds of *replication*. We might have replication in the sense of having carried out the experiment with more than 1 person in each cell (but with different persons from cell to cell) as in Table 47, or we might have a replication of measures on the same person or persons. Thus in Table 44 we could have  $m$  measures per person under each of the  $C$  conditions.

† Some of the confusion in textbooks (including the first edition of this one) regarding the choice of the error term is likely due to the fact that they were written before the models were explicitly stated. A complete statement of the models is given in E. G. Mentzer's 1953 pamphlet "Tests by the analysis of variance," prepared under the direction of Dr. Paul R. Rider and available as *WADC Technical Report 53-23*, Wright Air Development Center, Wright-Patterson Air Force Base, Ohio.

(We are not here concerned with replication in the sense of a repetition of the entire experiment by another investigator.)

Actually, for the working statistician the precise formula for the possible mathematical models is not nearly so important as the deductions therefrom regarding the meanings of the several variance estimates. Earlier (p. 253) we attempted to explain the meaning of the variance estimate,  $s^2_b$ . Perhaps the student should review the steps that led us to say that  $s^2_b$  is an estimate of  $\sigma^2 + m\sigma^2_{\bar{x}_m}$ . This we symbolized by an arrow, meaning "is estimate of." Another way of saying this is: the expected value of  $s^2_b$  is  $\sigma^2 + m\sigma^2_{\bar{x}_m}$ .

The general model for 2-way classification may be written as

$$(X_{rck} - \hat{\bar{X}}) = \alpha_r + \beta_c + (\alpha\beta)_{rc} + e_{rck} \quad (104)$$

in which the deviation of a score from the over-all population mean is thought of in terms of a row contribution,  $\alpha$ ; a column contribution,  $\beta$ ; an interaction effect,  $(\alpha\beta)$ ; and a normally distributed random error part,  $e_{rck}$ . The subscript  $k$  indicates that we have replication,  $m$  scores per cell, with  $k$  taking on values  $1 \cdots m$ , but the  $m$  scores in each cell are independent of the scores in all other cells. Both  $\alpha$  and  $\beta$ , and also  $(\alpha\beta)$  are expressed in deviation score form, i.e., possess the property that  $\sum_r \alpha_r = 0$ ,  $\sum_c \beta_c = 0$ , and  $\sum_r \sum_c (\alpha\beta)_{rc} = 0$ .

For the fixed constants model we replace  $\alpha$  and  $\beta$  by  $A$  and  $B$ , thus

$$(X_{rck} - \hat{\bar{X}}) = A_r + B_c + (AB)_{rc} + e_{rck} \quad (105a)$$

For the components of variance model we replace  $\alpha$  and  $\beta$  by  $a$  and  $b$ , thus

$$(X_{rck} - \hat{\bar{X}}) = a_r + b_c + (ab)_{rc} + e_{rck} \quad (105b)$$

and the mixed model can be written as (with columns standing for fixed constants)

$$(X_{rck} - \hat{\bar{X}}) = a_r + B_c + (aB)_{rc} + e_{rck} \quad (105c)$$

The  $a_r$ ,  $b_c$ ,  $(ab)_{rc}$ , and  $(aB)_{rc}$  are all assumed to be random samples from normally distributed populations of effects having variances



of  $\sigma_a^2$ ,  $\sigma_b^2$ ,  $\sigma_{ab}^2$ , and  $\sigma_{aB}^2$ . For the fixed values  $A_r$ ,  $B_c$ , and  $(AB)_{rc}$ , no assumption as to distribution of effects is required.

When the  $m$  scores per cell represent measurement replication,  $s_w^2$  will be taken as an estimate of  $\sigma_e^2$ ; when the  $m$  scores per cell involve  $m$  individuals,  $s_w^2$  will be regarded as an estimate of individual difference variance, designated by  $\sigma_i^2$ . It is to be understood that  $\sigma_i^2$  has 2 components: true score variance and error of measurement variance.

We are now ready to examine the various possible situations involving 2-way classification in order to point out just what is being estimated by  $s_r^2$ ,  $s_c^2$ ,  $s_{rc}^2$ , and  $s_w^2$ . Once this is done, we will be in a position to choose an appropriate variance estimate as the denominator, or error, term for  $F$ . The question of variance homogeneity will be discussed after a consideration of 9 situations (cases) involving 2-way classification (p. 311).

**Case I.** Fixed constants model, with  $m$  scores ( $m$  persons) per cell, a total of  $mRC$  individuals:

$$s_r^2 \rightarrow \sigma_i^2 + \frac{mC}{R-1} \sum A_r^2$$

$$s_c^2 \rightarrow \sigma_i^2 + \frac{mR}{C-1} \sum B_c^2$$

$$s_{rc}^2 \rightarrow \sigma_i^2 + \frac{m}{(R-1)(C-1)} \sum \sum (AB)_{rc}^2$$

$$s_w^2 \rightarrow \sigma_i^2$$

The general principle in forming an  $F$  ratio is to choose 2 estimates which differ (in their expected values) by 1 term only, the term involving the effect being tested. Accordingly,  $s_w^2$  is the correct denominator for  $F_r$ ,  $F_c$ , and  $F_{rc}$ , for testing row, column, and interaction effects, respectively. Note that interaction, if present, has nothing whatsoever to do with the main (row and column) effects. This is true because the interaction is a fixed, not a random, effect. If the interaction is significant, one must be on guard in drawing conclusions about the main effects—qualifications will be needed, as we learned in our discussion (pp. 301-303) of the meaning of interaction.

**Case II.** Fixed constants model,  $RC$  individuals, 1 in each cell: For this situation we have no  $s_w^2$ , hence no estimate of  $\sigma_i^2$ , but

the other estimates have precisely the same expected values here as under Case I, with  $m = 1$ . It is readily seen that we cannot form any  $F$  ratios for this situation—no significance test is possible; hence such an experimental setup should be avoided. If one can make the a priori assumption of zero interaction, one can use  $s^2_{rc}$  as the error term for  $F_r$  and  $F_c$ . That such an assumption may be indefensible is indicated by the fact that significant interactions have emerged in about half of the psychological research studies where interactions were testable. Note, however, that if the use of  $s^2_{rc}$  leads to a significant  $F$  for either rows or for columns, significance can be safely claimed since the used error term will tend to be too large because of interaction. The real danger is that the use of  $s^2_{rc}$  will too often lead to a false acceptance of the null hypothesis, and hence the overlooking of a real effect.

**Case III.** Fixed constants model,  $RC$  individuals, 1 per cell but each is measured  $m$  times:

$$s^2_r \rightarrow \sigma^2_i + \frac{mC}{R-1} \sum_r A^2_r$$

$$s^2_c \rightarrow \sigma^2_i + \frac{mR}{C-1} \sum_c B^2_c$$

$$s^2_{rc} \rightarrow \sigma^2_i + \frac{m}{(R-1)(C-1)} \sum_r \sum_c (AB)^2_{rc}$$

$$s^2_w \rightarrow \sigma^2_e$$

We see immediately that this design has exactly the same difficulties as Case II. The resulting  $s^2_w$  estimate is useless; if we did use  $s^2_w$  as the denominator for testing, say  $s^2_r$ , a significant  $F$  would be meaningless because we would not know whether its significance was attributable to a real row effect or to real individual differences or to a combination of the 2.

**Case IV.** Fixed constants model, only 1 person measured  $m$  times under each of the  $RC$  conditions: if we replace  $\sigma^2_i$  by  $\sigma^2_e$  in the last set of expected values we will have indicated what each  $s^2$  estimates. As for Case I, the appropriate error term for all 3  $F$ 's is  $s^2_w$ , but any conclusion one draws from a significant  $F$  must be carefully scrutinized for meaning. It can only mean

that the effect holds for the 1 person used in the experiment, with no assurance whatsoever that a repetition of the experiment with another person, either in the same or in a different laboratory, will lead to a confirmation of the results. In other words no generalization is possible except the trivial one that the effect holds for a particular individual, useful only in case one's scientific horizon is limited to 1 person.

The foregoing cases just about exhaust the possible situations for 2-way classification involving the fixed constants model. If it has occurred to the reader that each of  $m$  cases might be measured under all the  $RC$  conditions, he should be apprised that this would involve 3-way classification, to be discussed later. The important thing to have noted is that clear-cut results, permitting generalizations to a population of individuals, are possible only by the setup of Case I. We have listed the other 3 cases because it may be helpful to know what not to do.

**Case V.** Components of variance model, rows stand for  $R$  individuals and columns stand for, say,  $C$  judges, with  $m$  (ordinarily  $m$  will not exceed 2) ratings by each judge on each individual. The ratings, which must all be directed toward the same trait, might be based on observed, or on a transcribed record of, behavior of the  $R$  individuals. (The judges might find it difficult to rule out memory when making the ratings.) Instead of  $C$  judges making ratings we might have  $C$  examiners or testers, each testing each of the  $R$  individuals twice on, say, the Rorschach. We have a sample of individuals and a sample of judges (or examiners). The expected values of the variance estimates are:

$$s^2_r \rightarrow \sigma^2_e + m\sigma^2_{ab} + mC\sigma^2_a$$

$$s^2_c \rightarrow \sigma^2_e + m\sigma^2_{ab} + mR\sigma^2_b$$

$$s^2_{rc} \rightarrow \sigma^2_e + m\sigma^2_{ab}$$

$$s^2_w \rightarrow \sigma^2_e$$

It is obvious that  $s^2_w$  can be used as the error term for testing the interactive effect, but since  $s^2_w$  is nothing more than an estimate of error of measurement variance, the conclusion from a significant  $F$  is that interaction holds only for these particular  $R$  individuals and  $C$  judges—no assurance that repetition of the investigation with  $R$  other individuals and  $C$  other judges would

lead to interaction. As to the main effects, it is obvious that  $s^2_{rc}$  becomes the appropriate (and only correct) term to use for  $F_r$  and  $F_c$ . A significant  $F_r$  would mean a dependable differentiation of individual variation over and above variation due to measurement error and judge by individual interaction, and a significant  $F_c$  would indicate real variation from judge to judge in a possible population of judges.

**Case VI.** Components of variance model, same as Case V except that  $m = 1$ . No estimate of  $\sigma^2_e$  is available, but  $s^2_{rc}$  would still be the error term for both  $F$ 's.

Remark on components of variance model: Actually one is hard put to find good illustrations in psychology for this model. Aside from the illustration given above, it is difficult to find other classifications based on sampling. Any student who attempts to find other illustrations should keep in mind that it must be possible to classify a given score simultaneously in 2 different ways, each involving sampling.

**Case VII.** Mixed model, rows stand for  $R$  individuals, columns involve  $C$  fixed constants (fixed conditions having fixed effects), and measurement replication leading to  $m$  scores per cell:

$$s^2_r \rightarrow \sigma^2_e + m\sigma^2_{aB} + mC\sigma^2_a$$

$$s^2_c \rightarrow \sigma^2_e + m\sigma^2_{aB} + \frac{mR}{C-1} \sum B^2_c$$

$$s^2_{rc} \rightarrow \sigma^2_e + m\sigma^2_{aB}$$

$$s^2_w \rightarrow \sigma^2_e$$

The interaction term can be tested by  $F_{rc} = s^2_{rc}/s^2_w$ ; if  $F_{rc}$  is significant one concludes that the differential responses shown by these  $R$  individuals are larger than expected on the basis of error of measurement. Individual by conditions interactions are usually found to be significant. It will be recalled that in the mixed model the interaction term,  $(aB)_{rc}$ , is regarded as a random variable, and as such it becomes a source of random variation which, if real, will affect both main effects, both the between-rows and the between-columns terms. We see from the foregoing that  $s^2_{rc}$  becomes the proper error term for testing both  $s^2_r$  and  $s^2_c$ . To use  $s^2_w$  for this purpose is simply not defensible; if, for example,  $s^2_c/s^2_w$  is significant it might be so because of real column differ-

ences or because of real interaction or because of a combination of the 2. Ordinarily  $s^2_r$  in this situation is not tested for significance since it reflects individual differences which are always real unless the measurements are completely unreliable.

**Case VIII.** Mixed model, same as Case VII except that  $m = 1$  (no measurement replication). This does not provide an  $s^2_w$ , but  $s^2_{rc}$  is again the error term for testing both  $s^2_r$  and  $s^2_c$ . The setup in Table 44 falls under Case VIII.

**Case IX.** Mixed model,  $R$  rows stand for  $R$  individuals and columns stand for  $C$  forms of a test (the reliability of measurement setup discussed on pp. 290-294):

$$s^2_r \rightarrow \sigma^2_e + mC\sigma^2_a$$

$$s^2_c \rightarrow \sigma^2_e + \frac{mR}{C-1} \sum B_c^2$$

$$s^2_{rc} \rightarrow \sigma^2_e$$

It will be recalled that  $s^2_{rc}$ , which was previously (p. 287) labeled a remainder term, was shown to depend solely upon errors of measurement under the assumptions usually made in connection with test reliability. We see now that these assumptions involve the a priori assumption of no interaction, an assumption which implies, among other things, that possible practice effects are not differential from person to person. Note that in case interaction is operating, all 3 of the variance estimates in Case IX will involve interaction in the manner indicated for Case VII; hence  $s^2_{rc}$  is the appropriate error term regardless of whether there is or is not interaction.  $F_c = s^2_c/s^2_{rc}$  would be a test of the difference between form means or over-all practice effect or both (one wouldn't know which), and  $s^2_r/s^2_{rc}$  would be a test of whether reliable discriminations between individuals were being made, in spite of interaction if present.

Remark about measurement replication: We have seen that having  $s^2_w$  as an estimate of  $\sigma^2_e$  does not provide us with a useful error term (for  $F$ ) in the testing of hypotheses about main effects (and sometimes about interaction) under any of the 3 mathematical models. This illustrates a general principle: when an estimate of error of measurement variance is used as the denominator of  $F$ , no generalization to a population of persons is possible, and



hence no generalization of import to science. This raises the question as to whether measurement replication is worth while. The answer is yes, particularly when it is known that a single measurement is not very reliable. By replicating measurement we will obtain more reliable scores in the form of the average of  $m$  values; hence one source of variability in the data will be reduced. The student who has not noticed that the analyses involving measurement replication are, in essence, dealing with average scores for individuals should ponder further.

**Homogeneity of variance assumption.** For Cases I, II, and III it is assumed that individual difference variance is the same from cell to cell. For Cases IV through IX it is assumed that error of measurement variance is homogeneous from cell to cell. The assumption is testable (say, by Bartlett's test, p. 248) only for Cases I, IV, V, and VII.

### TRIPLE CLASSIFICATION

Suppose that we wish to arrange an investigation so as to let one set of data serve to determine whether the variation of a dependent variable is due to or associated with variation on 3 independent variables. Again, the term independent variable is being used in its broad sense. It might be a "real" variable like illumination, temperature, amount of food, length of rest interval; or it might be a variable having to do with qualitative differences, such as kind of food, type of motivation or incentive, various psychological sets. It makes no difference whether the variables are manipulatable in the laboratory, as would be true of all those mentioned, or whether the desired variation is secured by appropriate choice of cases.

It is necessary that we be able to assign individuals or scores to each combination of groupings made possible by whatever classifications we have on the 3 independent variables. Let us suppose that there are  $C$  categories on one variable,  $R$  on another, and  $B$  on a third. For purposes of exposition and as a systematic way of arranging the data, let the  $C$  categories define  $C$  columns, the  $R$  categories  $R$  rows, and the  $B$  categories  $B$  blocks. Let  $X_{rbc}$  represent the score in the  $r$ th row,  $b$ th block, and  $c$ th column, and let us assume for the time being that we have only 1 score for each combination. Thus  $X_{324}$  would be the only score in the



Table 51. SCORE AND SUM SCHEMA FOR TRIPLE CLASSIFICATION

		Column			Sum	Mean
		1	c	C		
Block 1	Row 1	$X_{111}$	$X_{11c}$	$X_{11C}$	$\sum^c X_{11c}$	$\bar{X}_{11.}$
	r	$X_{r11}$	$X_{r1c}$	$X_{r1C}$	$\sum^c X_{r1c}$	$\bar{X}_{r1.}$
	R	$X_{R11}$	$X_{R1c}$	$X_{R1C}$	$\sum^c X_{R1c}$	$\bar{X}_{R1.}$
	Sum Mean	$\sum^r X_{r11}$ $\bar{X}_{.11}$	$\sum^r X_{r1c}$ $\bar{X}_{.1c}$	$\sum^r X_{r1C}$ $\bar{X}_{.1C}$	$\sum^r \sum^c X_{r1c}$ $\bar{X}_{.1.}$	$\bar{X}_{.1.}$ Mean block 1
Block b	1	$X_{1b1}$	$X_{1bc}$	$X_{1bC}$	$\sum^c X_{1bc}$	$\bar{X}_{1b.}$
	r	$X_{rb1}$	$X_{rbc}$	$X_{rbC}$	$\sum^c X_{rbc}$	$\bar{X}_{rb.}$
	R	$X_{Rb1}$	$X_{Rbc}$	$X_{RbC}$	$\sum^c X_{Rbc}$	$\bar{X}_{Rb.}$
	Sum Mean	$\sum^r X_{rb1}$ $\bar{X}_{.b1}$	$\sum^r X_{rbc}$ $\bar{X}_{.bc}$	$\sum^r X_{rbC}$ $\bar{X}_{.bC}$	$\sum^r \sum^c X_{rbc}$ $\bar{X}_{.b.}$	$\bar{X}_{.b.}$ Mean block b
Block B	1	$X_{1B1}$	$X_{1Bc}$	$X_{1BC}$	$\sum^c X_{1Bc}$	$\bar{X}_{1B.}$
	r	$X_{rB1}$	$X_{rBc}$	$X_{rBC}$	$\sum^c X_{rBc}$	$\bar{X}_{rB.}$
	R	$X_{RB1}$	$X_{RBc}$	$X_{RBC}$	$\sum^c X_{RBc}$	$\bar{X}_{RB.}$
	Sum Mean	$\sum^r X_{rB1}$ $\bar{X}_{.B1}$	$\sum^r X_{rBc}$ $\bar{X}_{.Bc}$	$\sum^r X_{rBC}$ $\bar{X}_{.BC}$	$\sum^r \sum^c X_{rBc}$ $\bar{X}_{.B.}$	$\bar{X}_{.B.}$ Mean block B
Sums through blocks	1	$\sum^b X_{1b1}$	$\sum^b X_{1bc}$	$\sum^b X_{1bC}$	$\sum^b \sum^c X_{1bc}$	$\bar{X}_{1..}$
	r	$\sum^b X_{rb1}$	$\sum^b X_{rbc}$	$\sum^b X_{rbC}$	$\sum^b \sum^c X_{rbc}$	$\bar{X}_{r..}$
	R	$\sum^b X_{Rb1}$	$\sum^b X_{Rbc}$	$\sum^b X_{RbC}$	$\sum^b \sum^c X_{Rbc}$	$\bar{X}_{R..}$
	Sum	$\sum^r \sum^b X_{rb1}$	$\sum^r \sum^b X_{rbc}$	$\sum^r \sum^b X_{rbC}$	$\sum^r \sum^b \sum^c X_{rbc}$	$\bar{X}_{...}$
Means for rows by columns	1	$\bar{X}_{1..1}$	$\bar{X}_{1..c}$	$\bar{X}_{1..C}$	$\bar{X}_{1..}$	Means for rows
	r	$\bar{X}_{r..1}$	$\bar{X}_{r..c}$	$\bar{X}_{r..C}$	$\bar{X}_{r..}$	
	R	$\bar{X}_{R..1}$	$\bar{X}_{R..c}$	$\bar{X}_{R..C}$	$\bar{X}_{R..}$	
Column means		$\bar{X}_{..1}$	$\bar{X}_{..c}$	$\bar{X}_{..C}$	$\bar{X}_{...}$	$= \bar{X}$

third row, second block, and fourth column. The scores may be arranged in some such systematic order as that in Table 51, which should be studied carefully by the reader.

Note in particular how the various sums are specified and their location in the table. The first 2 subscripts in  $\sum^c X_{11c}$  indicate that this sum has to do with scores in the first row and first block, and that in the summing process  $c$  takes on values running from 1 to  $C$ . The general expression for all such sums is  $\sum^c X_{rbc}$ . The symbol  $\sum^r X_{r11}$  stands for the sum of scores in the first column and first block;  $r$  takes on values of 1 to  $R$ . The corresponding general symbol is  $\sum^r X_{rbc}$ . In next to the bottom section of the table will be found  $\sum_b X_{1b1}$  as the sum for all the cases in row 1 and column 1, the summing being through blocks; i.e.,  $b$  takes on values from 1 to  $B$ . The general expression for such sums is  $\sum_b X_{rbc}$ . The sum of all the scores in the first block is symbolized as  $\sum_{r,c} X_{r1c}$ , and in the  $b$ th block as  $\sum_{r,c} X_{rbc}$ . For the sum of all the scores in the first column, irrespective of row and block, we have  $\sum_{r,b} X_{r1b}$ , and the general expression is  $\sum_{r,b} X_{rbc}$ . The symbol  $\sum_{b,c} X_{1bc}$  stands for the sum of all scores in the first row, and  $\sum_{b,c} X_{rbc}$  is the corresponding general expression. Note also how the "dot" notation is used to specify the several means. The subscript which has been replaced by a dot indicates the direction of the addition required to obtain the sum for the given mean. Thus in  $\bar{X}_{.24}$  the dot replaces  $r$ ; this mean is based on  $R$  scores, with  $r$  running from 1 to  $R$  when we sum. The subscripts which are left denote that the mean is for scores in the second block and fourth column. The total number of means will be as follows:

$RB$  means of the form  $\bar{X}_{rb}$ .

$RC$  means of the form  $\bar{X}_{r.c}$ .

$BC$  means of the form  $\bar{X}_{.bc}$ .

$R$  means of the form  $\bar{X}_{r..}$ .

$B$  means of the form  $\bar{X}_{.b.}$ .

$C$  means of the form  $\bar{X}_{..c}$ .

One mean of the form  $\bar{X}_{...}$  = total mean =  $\bar{X}$

Perhaps a better appreciation of the meaning of all these means can be obtained by a study of Fig. 19, which pictures geometrically the situation for 2 blocks, 3 rows, and 4 columns. The individual scores can be thought of as in the cubicles of a 2 by 3 by 4 box. Summing through the box in the vertical direction leads to the

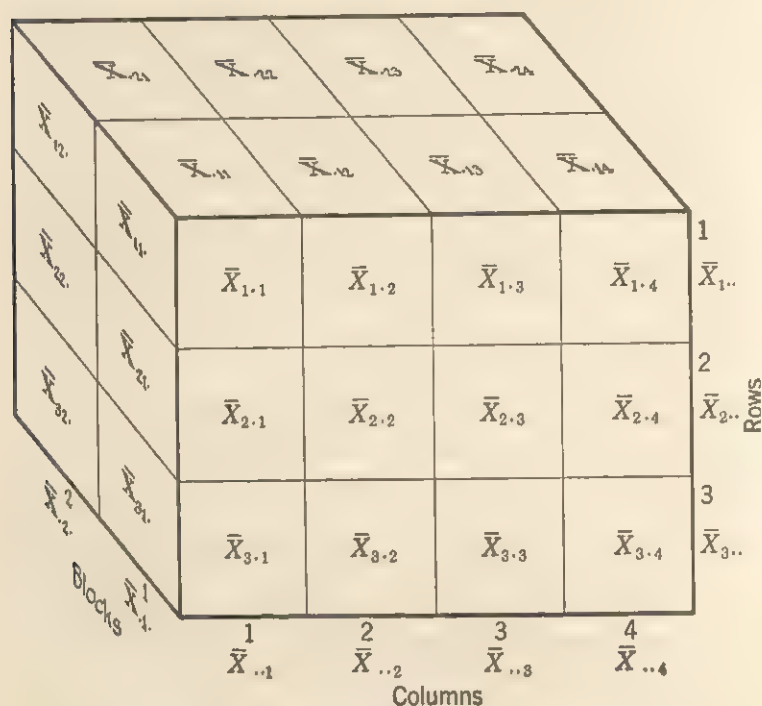


Fig. 19. Geometric picture of 3-way classification.

8 means on the top; summing in the forward-backward direction leads to the 12 means on the front surface; and summing through right-leftward leads to the 6 means on the side. Summing the means (or summing sums) across the front leads to the means placed along the vertical axis for the groups defined by the rows; summing the means (or sums) downward on the front leads to the means placed along the right-left axis for the groups defined by the columns; summing down on the side leads to the means along the third axis for the groups defined by the blocks. To

get any of these means it is, of course, assumed that the sum involved is divided by the proper number.

Of primary interest is the question: Is the variation among the means along the edges, considered separately, larger than expected on the basis of chance? To answer this we need to break down the sum of squares of deviations from the total mean into appropriate components. The score  $X_{rbc}$  in the cubicle defined by the  $r$ th row,  $b$ th block, and  $c$ th column will vary more or less from  $\bar{X}$ , and 3 possible sources of variation for  $X_{rbc}$  are obvious: the deviation of its row mean, its column mean, and its block mean from  $\bar{X}$ . Now, if we recall the situation for double classification, it is fairly obvious that, when the score  $X_{rbc}$  is considered as belonging in row  $r$  and column  $c$ , one source of variation becomes the remainder or interaction for rows and columns; considered next as also falling in row  $r$  and block  $b$ , another source of variation is the possible interaction of rows and blocks; and then thought of as belonging to column  $c$  and block  $b$ , the score also involves the interaction of columns and blocks.

When the sums of squares for these 6 components are added, it will be discovered that they do not sum to the sum of squares for the total; i.e., subtracting these 6 sums from the total sum leaves a remainder. This residual is sometimes referred to as error, more frequently as a *triple interaction*. This term involves rows, blocks, and columns. The reader, having in mind the idea that the simple row by column interaction has to do with the possible failure of cell entries to be consistent with the 2 sets of marginal means, must now try imagining that the  $RBC$  entries in the cubical cells of our box may not be entirely consistent with the 3 sets of means on the edges and with the 3 sets on the surface. We have seen that a statistical check on simple interaction is not possible with only 1 entry per cell; similarly more than 1 score per cubicle is required for testing triple interaction.

Table 52 gives the essentials, in symbols, for the analysis of variance for the triple classification setup. In order to specify the interactions, we here adopt the abbreviation scheme generally used. Thus  $R \times B$ , read  $R$  by  $B$ , indicates the row and block interaction, and  $R \times B \times C$  stands for the row by block by column or triple interaction. In a given investigation, the rows, blocks, and columns refer to particular independent or classificatory variables.

Table 52 VARIANCE TABLE FOR TRIPLE CLASSIFICATION INTO  $R$  ROWS,  $B$  BLOCKS, AND  $C$  COLUMNS

Source	Sum of Squares	$df$	Variance Estimate
Rows	$BC\sum^r(\bar{X}_{r..} - \bar{X})^2$	$R-1$	$s^2_r$
Blocks	$RC\sum^b(\bar{X}_{.b.} - \bar{X})^2$	$B-1$	$s^2_b$
Columns	$RB\sum^c(\bar{X}_{..c} - \bar{X})^2$	$C-1$	$s^2_c$
$R \times B$ interaction	$C\sum^{rb}(\bar{X}_{rb.} - \bar{X}_{r..} - \bar{X}_{.b.} + \bar{X})^2$	$(R-1)(B-1)$	$s^2_{rb}$
$R \times C$ interaction	$B\sum^{rc}(\bar{X}_{r.c.} - \bar{X}_{r..} - \bar{X}_{..c} + \bar{X})^2$	$(R-1)(C-1)$	$s^2_{rc}$
$B \times C$ interaction	$R\sum^{bc}(\bar{X}_{.bc.} - \bar{X}_{.b.} - \bar{X}_{..c} + \bar{X})^2$	$(B-1)(C-1)$	$s^2_{bc}$
$R \times B \times C$ or triple interaction	$\sum^{rbc}(\bar{X}_{rbc.} - \bar{X}_{rb.} - \bar{X}_{r.c.} - \bar{X}_{.bc.} + \bar{X}_{r..} + \bar{X}_{.b.} + \bar{X}_{..c} - \bar{X})^2$	$(R-1)(B-1)(C-1)$	$s^2_{rbc}$
Total	$\sum^{rbc}(X_{rbc} - \bar{X})^2$	$RBC-1$	

It will be noted in Table 52 that the  $df$  for the triple interaction term is given as  $(R-1)(B-1)(C-1)$ . The student may be helped in understanding the reasoning which leads to this  $df$  by referring again to Fig. 19. The surface means tend to restrict the deviation score values within the box. How many cubical cells can we fill before these restrictions operate? The general rule-of-thumb procedure for determining the  $df$  for interaction sums of squares is to take the product of the  $df$ 's of the variables involved in the given interaction. This holds for simple, triple, and higher-order interactions.

#### SPECIAL CASE WHERE THE ROWS STAND FOR PERSONS OR MATCHED INDIVIDUALS

Suppose the purpose of a study is to ascertain whether variation on a dependent variable is influenced by or associated with variation on 2 independent variables. This, of course, involves

the double classification idea previously discussed, but we are now in a position to accomplish, by means of triple classification, 2 closely related things which could not be done by the simpler double classification scheme.

1. If transfer, practice, fatigue, etc., effects are such that it is permissible to make observations on an individual under each of the  $RC$  combinations of conditions, we may increase the precision of an experiment by using only  $m$  individuals instead of  $mRC$  individuals, as in the illustration involving pursuit learning. Or we may make observations on  $mRC$  cases so as to have in each of the  $RC$  cells  $m$  scores which are based on  $m$  sets of matched individuals, thereby reducing error.

2. If we are dealing with a situation in which it is required that observations be made on the same individual in each of the  $RC$  conditions, and if more than 1 case is used either to reduce errors or to provide a basis for generalizing to a population, it is necessary that we make statistical allowance for the fact that the  $RC$  observations on the  $m$  cases are nonindependent, or correlated. This allowance was not possible by the double classification scheme, for which it was assumed that the  $m$  scores in 1 cell were independent of the observations in the other cells.

It will be recalled that in the double classification setup, by letting 1 classification refer to  $R$  individuals or sets of matched cases, we were provided with an over-all test of significance for several correlated means for groups classified on a *single* independent variable. Triple classification permits a similar test of correlated means for groups involved in *double* classification.

Since the assigning of the bases of classification to rows, blocks, and columns is arbitrary, we shall let the  $R$  rows stand for  $R$  individuals (or  $R$  matched persons), with the blocks and columns representing the independent variables to be investigated.

### COMPUTATIONAL ILLUSTRATION FOR TRIPLE CLASSIFICATION

The task of computing the required sums of squares (see Table 52) is tedious. The first step is to arrange the data in some such systematic order as that depicted in Table 51 and do the necessary adding to secure the various sums indicated in that table. The total sum of squares for all  $RBC$  cases is obtained as usual:



sum all the scores, sum all the squared scores, and substitute in the general formula  $(1/RBC)[RB(\sum X^2 - (\sum X)^2)]$ .

To secure the 3 between-groups and the 3 simple interaction sums of squares, we form 3 subtables involving sums taken in various directions. For the first of these subtables we take row by column sums obtained by adding cell entries from block to block, i.e., through the  $B$  blocks. The next to the bottom section of Table 51 contains these row by column sums, which we reproduce here as Table 53a. The reader will note that the values for Table 53b are the right-hand margin sums of Table 51, and that the values for Table 53c are found as the sums in Table 51 along the bottom of each block.

With these auxiliary tables in mind, we can write the required computational formulas. The simple interaction terms are secured by computing a subtotal sum of squares for each table and then subtracting therefrom the 2 appropriate "between" sums of squares. These subtotal sums of squares will not be the same as the total sum of squares obtained for double classification by formula (102a) because we are now dealing with cell entries which are the sums of scores rather than single scores. Due allowance for this can be made by a slight change in formula (102a). The amended formula, with notation appropriate for and specific to the 3 auxiliary tables, may be written as follows:

Subtotal: row by column

$$\frac{1}{RBC} [RC\sum\sum(\sum X_{rbc})^2 - (\sum\sum\sum X_{rbc})^2] \quad (106a)$$

Subtotal: row by block

$$\frac{1}{RBC} [RB\sum\sum(\sum X_{rbc})^2 - (\sum\sum\sum X_{rbc})^2] \quad (106b)$$

Subtotal: block by column

$$\frac{1}{RBC} [BC\sum\sum(\sum X_{rbc})^2 - (\sum\sum\sum X_{rbc})^2] \quad (106c)$$

From the right-hand margin of either Table 53a or 53b we can compute the sum of squares for

$$\text{Between rows: } \frac{1}{RBC} [R\sum(\sum\sum X_{rbc})^2 - (\sum\sum\sum X_{rbc})^2] \quad (106d)$$

Table 53a. REQUIRED SUMS FOR ROW BY COLUMN ANALYSIS

	1	c	C	Sum
1	$\sum X_{1b1}$	$\sum X_{1bc}$	$\sum X_{1bC}$	$\sum \sum X_{1bc}$
r	$\sum X_{rb1}$	$\sum X_{rbc}$	$\sum X_{r b C}$	$\sum \sum X_{rbc}$
R	$\sum X_{Rb1}$	$\sum X_{Rbc}$	$\sum X_{R b C}$	$\sum \sum X_{Rbc}$
Sum	$\sum \sum X_{rb1}$	$\sum \sum X_{rbc}$	$\sum \sum X_{r b C}$	$\sum \sum \sum X_{rbc}$

Table 53b. REQUIRED SUMS FOR ROW BY BLOCK ANALYSIS

	1	b	B	Sum
1	$\sum X_{11c}$	$\sum X_{1bc}$	$\sum X_{1Bc}$	$\sum \sum X_{1bc}$
r	$\sum X_{r1c}$	$\sum X_{rbc}$	$\sum X_{rBc}$	$\sum \sum X_{rbc}$
R	$\sum X_{R1c}$	$\sum X_{Rbc}$	$\sum X_{RBc}$	$\sum \sum X_{Rbc}$
Sum	$\sum \sum X_{r1c}$	$\sum \sum X_{rbc}$	$\sum \sum X_{rBc}$	$\sum \sum \sum X_{rbc}$

Table 53c. REQUIRED SUMS FOR BLOCK BY COLUMN ANALYSIS

	1	c	C	Sum
1	$\sum X_{r11}$	$\sum X_{r1c}$	$\sum X_{r1C}$	$\sum \sum X_{r1c}$
b	$\sum X_{rb1}$	$\sum X_{rbc}$	$\sum X_{r b C}$	$\sum \sum X_{rbc}$
B	$\sum X_{rB1}$	$\sum X_{rBc}$	$\sum X_{r B C}$	$\sum \sum X_{r B c}$
Sum	$\sum \sum X_{rb1}$	$\sum \sum X_{rbc}$	$\sum \sum X_{r b C}$	$\sum \sum \sum X_{rbc}$

From the bottom of either Table 53a or 53c we can obtain the sum of squares for

$$\text{Between columns: } \frac{1}{RBC} [C\sum^c(\sum^r\sum^bX_{rbc})^2 - (\sum^r\sum^b\sum^cX_{rbc})^2] \quad (106e)$$

From the bottom of Table 53b or from the right-hand margin of 53c we can calculate the sum of squares for

$$\text{Between blocks: } \frac{1}{RBC} [B\sum^b(\sum^r\sum^cX_{rbc})^2 - (\sum^r\sum^b\sum^cX_{rbc})^2] \quad (106f)$$

Then from the above 6 sums of squares the simple interaction sums of squares may be secured by the following subtractions:

$$\text{Row by column interaction: } (106a) - (106d) - (106e) \quad (107a)$$

$$\text{Row by block interaction: } (106b) - (106d) - (106f) \quad (107b)$$

$$\text{Block by column interaction: } (106c) - (106e) - (106f) \quad (107c)$$

And finally, again by subtraction, we have the sum of squares for the row by column by block, or

$$\text{Triple interaction: Total sum of squares minus (106def) minus (107abc).}$$

We will illustrate the procedure by using the data of Table 54, in which the blocks represent 2 levels of illumination, the columns 3 degrees of albedo, and the rows 4 individuals, and the scores are judged whiteness. Notice that each subject made judgments under all 6 of the combinations of conditions. The sums given in Table 54 become the entries for the auxiliary computational Tables 55abc. The needed value of  $\sum^r\sum^b\sum^cX_{rbc}$  is 898, and the sum of all the squared scores,  $\sum^r\sum^b\sum^cX_{rbc}^2$ , is 44,394. From these figures we have

$$\frac{1}{24} [24(44,394) - (898)^2] = 10,793.83 = \text{total sum of squares}$$

The various "between" sums can readily be obtained by adding the squares of the appropriate marginal sums of auxiliary Tables 55abc, and substituting in formulas (106def).

$$\text{For between blocks we need } (414)^2 + (484)^2 = 405,652;$$

$$\text{For between columns we need } (152)^2 + (247)^2 + (499)^2 = 333,114;$$

# Computational Illustration for Triple Classification 321

For between rows we need  $(198)^2 + (202)^2 + (197)^2 + (301)^2 = 209,418$ .

Table 54. DATA USED IN ILLUSTRATING COMPUTATIONS FOR 3-WAY CLASSIFICATION: 2 LEVELS OF ILLUMINATION (BLOCKS), 3 ALBEDOS (COLUMNS), AND 4 OBSERVERS (ROWS) \*

Illumination	Observer	Albedo			Sum	Mean
		.07	.14	.26		
1.20	1	11	24	60	95	31.67
	2	22	26	44	92	30.67
	3	16	22	55	93	31.00
	4	20	32	82	134	44.67
	Sum	69	104	241	414	34.50
	Mean	17.25	26.00	60.25	34.50	
2.00	1	14	24	65	103	34.33
	2	27	36	47	110	36.67
	3	18	24	62	104	34.67
	4	24	59	84	167	55.67
	Sum	83	143	258	484	40.33
	Mean	20.75	35.75	64.50	40.33	
Sums through blocks	1	25	48	125	198	33.00
	2	49	62	91	202	33.67
	3	34	46	117	197	32.83
	4	44	91	166	301	50.17
	Sum	152	247	499	898	37.42
Means for rows by columns	1	12.50	24.00	62.50	33.00	
	2	24.50	31.00	45.50	33.67	
	3	17.00	23.00	58.50	32.83	
	4	22.00	45.50	83.00	50.17	
Column means		19.00	30.87	62.38	37.42	

\* Data from R. E. Taubman, *J. Exp. Psychol.*, 1945, 35, 235-241.

## Analysis of Variance: Complex

Table 55a. REQUIRED SUMS FOR BLOCK BY COLUMN ANALYSIS

Illumination	Albedo			Sum
	.07	.14	.26	
1.20	69	104	241	414
2.00	83	143	258	484
Sum	152	247	499	898

Table 55b. REQUIRED SUMS FOR ROW BY BLOCK ANALYSIS

Illumination	Individuals				Sum
	1	2	3	4	
1.20	95	92	93	134	414
2.00	103	110	104	167	484
Sum	198	202	197	301	898

Table 55c. REQUIRED SUMS FOR ROW BY COLUMN ANALYSIS

Individual	Albedo			Sum
	.07	.14	.26	
1	25	48	125	198
2	49	62	91	202
3	34	46	117	197
4	44	91	166	301
Sum	152	247	499	898

Then we have

$$\frac{1}{24}[2(405,652) - (898)^2] = 204.17 \text{ for between-blocks sum of squares}$$

$$\frac{1}{24}[3(333,114) - (898)^2] = 8039.08 \text{ for between-columns sum of squares}$$

$$\frac{1}{24}[4(209,418) - (898)^2] = 1302.83 \text{ for between-rows sum of squares}$$

In order to secure the subtotal sums of squares we add the squares of the cell entries in the auxiliary tables. For the block by column subtotal we have from Table 55a:

$$(69)^2 + (83)^2 + (104)^2 + (143)^2 + (241)^2 + (258)^2 = 167,560$$

Similarly for the row by block subtotal we have from Table 55b:

$$(95)^2 + (103)^2 + \dots + (167)^2 = 105,508$$

and for the row by column subtotal we have from Table 55c:

$$(25)^2 + \dots + (44)^2 + \dots + (166)^2 = 87,814$$

These 3 sums can now be substituted into formulas (106abc):

$$\frac{1}{24}[6(167,560) - (898)^2] = 8289.83 = \text{block by column subtotal sum of squares}$$

$$\frac{1}{24}[8(105,508) - (898)^2] = 1569.17 = \text{row by block subtotal sum of squares}$$

$$\frac{1}{24}[12(87,814) - (898)^2] = 10,306.83 = \text{row by column subtotal sum of squares}$$

Next we get the simple interaction sum of squares by the subtractions indicated in formulas (107abc):

$$8289.83 - 204.17 - 8039.08 = 46.58 = \text{block by column interaction}$$

$$1569.17 - 204.17 - 1302.83 = 62.17 = \text{row by block interaction}$$

$$10,306.83 - 8039.08 - 1302.83 = 964.92 = \text{row by column interaction}$$



Then for the triple interaction sum of squares we have

$$10,793.83 - 204.17 - 8039.08 - 1302.83 \\ - 46.58 - 62.17 - 964.92 = 174.08$$

The several sums of squares, their *df*'s, and the resulting variance estimates are brought together in Table 56.

Table 56. ANALYSIS OF VARIANCE FOR JUDGED WHITENESS BY 4 OBSERVERS FOR 3 DEGREES OF ALBEDO AND 2 LEVELS OF ILLUMINATION

Source	Sum of Squares	<i>df</i>	Variance Estimate
Illumination	204.17	1	204.17
Albedo	8,039.08	2	4,019.54
Subjects (individual differences)	1,302.83	3	434.28
Interaction: $I \times A$	46.58	2	23.29
Interaction: $I \times S$	62.17	3	20.72
Interaction: $A \times S$	964.92	6	160.82
Interaction—triple: $I \times A \times S$	174.08	6	29.01
Total	10,793.83	23	

We are not yet ready to discuss the principles controlling the choice of the error term appropriate for the possible *F*'s. When the models have been presented, the student may check back to see whether we have used, in the next 2 paragraphs, the correct denominator for the *F* ratio.

First we use the triple interaction as a basis for testing the significance of the simple interactions. Of chief interest in this example is the possible interaction between albedo and illumination, but since this interaction variance is less than that for triple interaction, we know at once without computing *F* that the interaction is insignificant. The illumination by individual interaction is also insignificant. The interaction of albedo with individuals yields an *F* of  $160.82/29.01 = 5.54$ , which, for  $n_1 = 6$  and  $n_2 = 6$ , falls between the values of 4.28 and 8.47 for the .05 and .01 levels respectively. This *F* of 5.54 is high enough to suggest that the form of the relationship between judged whiteness and albedo varies somewhat from person to person.

Now we turn to a test of the main effects. A test of the significance of row differences is a test of individual differences and is accordingly of little interest. For illumination we have  $F = 204.17/20.72 = 9.85$ , which falls beyond the 5.99 required for  $P = .05$ , and is therefore suggestive of a real difference due to illumination. For albedo we have  $F = 4019.51/160.82 = 24.99$ , which is highly significant.

Actually, the foregoing results are not to be regarded as conclusive. The data which we have used to illustrate the computations are only a part of more complete data which involved additional degrees of albedo and other levels of illumination. Partly because of space limitations and partly because it is easier to illustrate the computations when only a few rows, columns, and blocks are involved, we have ignored a part of the available data.

It should be kept in mind that this illustration is an example of the use of the triple classification scheme as a method for making allowance for the use of correlated observations in a problem of double classification involving the influence of 2 variables on a third. In this special use of triple classification, in which the rows correspond to individuals, the objective is identical with that in the earlier analysis of pursuit rotor learning (Table 49). The 2 situations are similar in that there are  $m$  (or  $R$ ) scores in each cell; they are different in that the  $m$  scores in any one cell for the pursuit learning problem are independent of the  $m$  scores in other cells, whereas the  $R$  scores in each of the albedo-illumination cells are correlated—each person contributes a score to each cell. Both schemes permit a check on the interaction effect of the 2 independent variables used to classify the observations. The use of  $BC$  observations on each of  $R$  cases (if feasible) will yield more precise information than obtainable by having scores for  $m$  individuals in each of the  $BC$  cells. This is analogous to the well-known principle that experimentation in which individuals serve as their own controls tends to be more precise than that in which an independent control group is set up.

### TRIPLE CLASSIFICATION WITH $m$ CASES PER CUBICLE

We have seen how the possible association of a dependent variable with 3 independent variables can be tested by a variance analysis made on a triple classification basis. If one wishes either to base his results on more than  $RBC$  observations or to test the

significance of the triple interaction, it is necessary to have more than 1 score in each cubicle. This can be accomplished either by assigning  $m$  individuals to each of the  $RBC$  combinations of conditions or by using just  $m$  individuals with each yielding an observation under all the  $RBC$  conditions or by using  $m$  sets of  $RBC$  cases with 1 individual of each set assigned to each of the  $RBC$  groups. Matching may not be feasible; neither may the securing of  $RBC$  observations on each of  $m$  individuals be feasible. At times, however, the problem under consideration may require an observation on each individual under all the conditions. Whether  $m$  individuals are so used by preference or by necessity, we will have  $m$  measurements in each of the  $RBC$  cubicles, but in testing the significance of the differences between the means of rows or of columns or of blocks we will be dealing with a situation in which the means are correlated because they are based upon the *same* individuals. To allow for this fact we would need a quadruple classification setup.

Let us next consider the case in which we have in each cubicle  $m$  scores, which are independent of the  $m$  scores in other cubicles. The total number of scores will, of course, be  $mRBC$ , and the breakdown of the total sum of squares will include the components specified in Table 52 plus a within-cubicles sum of squares. Since each cubicle defines a group, the within-cubicles sum of squares does not differ from previously discussed "within" sums of squares. The formula in this case is

$$\frac{1}{m} [m \sum \sum \sum X^2_{rbc} - \sum (\sum X_{rbc})^2]$$

in which it is understood that the  $\sum X^2$  term contains  $mRBC$  squares and that the subtractive term indicates that we first sum the  $m$  scores separately for each cubicle, then square each of these sums, and finally sum all these  $RBC$  squared sums. The  $df$  for this term will be  $mRBC - RBC$  because we are dealing with the deviations of  $mRBC$  scores about  $RBC$  different means.

With  $m$  independent scores per cubicle, the 6 computational formulas (106) need only be modified by the use of  $1/mRBC$  instead of  $1/RBC$  as the factor outside the brackets. It must be understood, however, that the sums within the parentheses of formulas (106) will involve  $m$  times as many scores as for the simpler situation with 1 case per cubicle. The computation is

again accomplished by auxiliary tables, the main cell entries of which will, of course, also involve sums with  $m$  times as many scores. If we think of the orderly arrangement of the original data, as exemplified in Table 51, it will be seen that each cell in the separate block designations will consist of  $m$  score entries; i.e., we will have  $m$  scores of the type  $X_{111}$  or  $X_{324}$ . A more precise notation would be to let  $X_{irbc}$  stand for the score of the  $i$ th person in the  $r$ th row and  $c$ th column of the  $b$ th block, with  $i$  taking on values of 1, 2,  $\dots$   $m$ .

Except for the use of  $1/mRBC$  in place of  $1/RBC$  in formulas (106), the computation of the between and simple interaction sums of squares follows exactly the steps outlined for a single score per cubicle. The triple interaction sum of squares is again obtained by subtraction, *but* now we must also deduct the within-cubicles sum of squares. Note that in the formula of Table 52 which defines the triple interaction term we need to replace  $X_{rbc}$  by  $\bar{X}_{rbc}$ , the mean of the  $m$  scores in the  $r$ th row and  $c$ th column of block  $b$ .

### CHOICE OF ERROR TERM IN 3-WAY CLASSIFICATION

The general mathematical model for the breakdown of a score in the triple classification setup may be written as

$$(X_{rbck} - \hat{X}) \\ = \alpha_r + \delta_b + \gamma_c + (\alpha\delta)_{rb} + (\alpha\gamma)_{rc} + (\delta\gamma)_{bc} + (\alpha\delta\gamma)_{rbc} + e_{rbck}$$

in which the subscripts,  $r$ ,  $b$ , and  $c$  refer to rows, blocks, and columns, and  $k$  takes on values 1  $\dots$   $m$ , there being  $m$  independent replications (either of measurement or of individuals) in each cell. The mean value of each term on the right of the equality sign is zero; that is, all values are expressed in deviation units. Note the manner in which the interactive effects are designated. Using notation analogous to that employed in specifying equations (105) from equation (104) for 2-way classification, we may replace  $\alpha$ ,  $\delta$ , and  $\gamma$  by their Latin equivalents, with capitals  $A$ ,  $D$ , and  $G$  representing fixed values (fixed constants model), and with lower-case letters  $a$ ,  $d$ , and  $g$  standing for classifications involving samplings (components of variance model). The mixed model would, of

course, contain 1 of the lower case and 2 of the capital letters or 2 of the lower case and 1 of the capital letters.

Rather than rewrite the model equation with particular Latin letters specifying the particular models, we can indicate the models by the following symbols:

[ADG] for fixed constants model

[adg] for components of variance model

[aDG] and [adG] for mixed models

It is *assumed* that the  $a_r$ ,  $d_b$ ,  $g_c$ ,  $(ad)_{rb}$ ,  $(aD)_{rb}$ ,  $(ag)_{rc}$ ,  $(aG)_{rc}$ ,  $(dg)_{bc}$ ,  $(dG)_{bc}$ ,  $(adg)_{rbc}$ ,  $(adG)_{rbc}$ ,  $(aDG)_{rbc}$ , and  $e_{rbck}$  are random samples from normally distributed populations of effects having the respective variances:  $\sigma^2_a$ ,  $\sigma^2_d$ ,  $\sigma^2_g$ ,  $\sigma^2_{ad}$ ,  $\sigma^2_{aD}$ ,  $\sigma^2_{ag}$ ,  $\sigma^2_{aG}$ ,  $\sigma^2_{dg}$ ,  $\sigma^2_{dG}$ ,  $\sigma^2_{adg}$ ,  $\sigma^2_{adG}$ ,  $\sigma^2_{aDG}$ , and  $\sigma^2_e$  when  $k = 1 \cdots m$  represents measurement replication or  $\sigma^2_i$  when  $k = 1 \cdots m$  involves replication of individuals. Seldom does one have an opportunity to check on the normality of the several interactive effects—a fact which may be disturbing to the reader. No such assumptions are made regarding the effects  $A_r$ ,  $D_b$ ,  $G_c$ ,  $(AD)_{rb}$ ,  $(AG)_{rc}$ ,  $(DG)_{bc}$ , and  $(ADG)_{rbc}$ , which are associated with the fixed constants. Since all effects are expressed in terms of deviation units, the sum of each particular set of effects, such as  $a_r$  or  $A_r$  or  $(aD)_{rb}$  or  $(DG)_{bc}$ , is zero.

In order to choose the appropriate variance estimate for the denominator of  $F$  for a given significance test, we again need to indicate just what each possible variance estimate ( $s^2$ ) estimates. A summary statement will be given later regarding the assumption of homogeneity of variance for the several cases involving 3-way classification (p. 335).

**Case X.** Fixed constants model [ADG], with  $m$  different individuals in each of the  $RBC$  cubicles. This is a simple straightforward case in which  $s^2_w \rightarrow \sigma^2_i$ , and *all* the other 7  $s^2$  values are estimates of  $\sigma^2_i$  *plus* a single (possible) effect, the one to be tested. Examples:

$$s^2_r \rightarrow \sigma^2_i + \frac{mBC}{R-1} \sum^r A^2_r$$

and

$$s^2_{rb} \rightarrow \sigma^2_i + \frac{mC}{(R-1)(B-1)} \sum^r \sum^b (AD)^2_{rb}$$



Thus  $s^2_w$  is the proper error term for testing all 3 main effects, all three 2-way interactions, and the 3-way interaction. Generalizations are to the population(s) from which the  $mRBC$  persons were drawn, but conclusions regarding main effects, or factors (the  $\alpha$ ,  $\delta$ , and  $\gamma$  are often spoken of as *factors*), will need to be qualified in case a given factor is involved in a significant interaction.

**Case XI.** Fixed constants model [ADG], with 1 individual (measured once) in each of the  $RBC$  cubicles, a total of  $RBC$  persons. This design yields no  $s^2_w$  by which to estimate  $\sigma^2_i$ , which, as for Case X, is involved in the expected value of each of the other variance estimates; hence no tests of significance are possible unless one can make (and defend) the a priori assumption that the 3-way interaction is zero. If so,  $s^2_{rbc}$  would become the error term for testing the main and the 2-way interaction effects. A defensible a priori assumption that any 1 of the 2-way interactions is zero would also provide the desired estimate of  $\sigma^2_i$  for testing the main effects and the other interactions. We repeat what was said in the discussion of the error term for the double classification setup: significant interactions are so prevalent in psychology that the a priori assumption of a zero interaction needs to be backed up by very strong logic. It should be noted that if the use of  $s^2_{rbc}$  (without justification of the assumption of zero 3-way interaction) leads to a significant  $F$ , we can be sure of its significance since this as an error term will be too large in case of nonzero 3-way interaction (compare with analogue in double classification, p. 307). What about the risk of a type II error?

**Case XII.** Fixed constants model [ADG], 1 person per cubicle but each person is measured  $m$  times. This leads to an  $s^2_w$  which is an estimate of  $\sigma^2_e$  rather than the needed estimate of  $\sigma^2_i$ . Now it might be thought that this  $s^2_w$  could be used to test  $s^2_{rbc}$  for the presence of 3-way interaction, but note that since  $s^2_{rbc} \rightarrow \sigma^2_i +$

$$\frac{m}{(R-1)(B-1)(C-1)} \sum \sum \sum (ADG)^2_{rbc} \text{ and } s^2_w \rightarrow \sigma^2_e, \text{ the division of } s^2_{rbc} \text{ by } s^2_w \text{ leads to a noninterpretable } F \text{ (if significant)}$$

because one has no way of knowing whether the significance is due to individual differences or to 3-way interaction (remember that  $\sigma^2_i$  contains an error of measurement part). Stated differently, the  $s^2_{rbc}$  is an estimate in which error of measurement variance, true individual difference variance, and possible 3-way interaction effects are all *confounded*, a term used to indicate that a given setup



does not allow a disentangling of the sources of variation which enter into a particular estimate. By using the score in each cubicle as the average of the  $m$  measurements, one can handle Case XII in the manner indicated for Case XI. The same difficulties are encountered—the only advantage of Case XII over Case XI is that the scores, being averages, are more reliable.

**Case XIII.** Fixed constants model  $[ADG]$ , with only 1 person supplying all scores, or a score (or scores) under each of the  $RBC$  conditions. If we have  $m$  measures on the 1 person under each of the possible combinations of conditions,  $s^2_{\alpha} \rightarrow \sigma^2_e$  and each of the other 7 variance estimates has an expected value including  $\sigma^2_e$  plus an effect. A significant  $F$  with  $s^2_w$  as the error term permits only the conclusion that repetition of the experiment on this same person would be expected to yield similar results—a “generalization” which has no generality, and hence is worthless. If the 1 person provides only 1 score per cubicle, we won’t even have an estimate of  $\sigma^2_e$ ; hence we need to make a priori assumptions about interactions (as for Case XI) in order to “generalize” to this 1 person. Thus Case XIII as a possible experimental design holds no promise.

**Case XIV.** Mixed model  $[aDG]$ . Typically, this will involve  $R$  individuals assigned to the rows with each measured at least once under the  $BC$  conditions. We have (with no measurement replication):

$$s^2_r \rightarrow \sigma^2_e + \sigma^2_{aDG} + C\sigma^2_{aD} + B\sigma^2_{aG} + BC\sigma^2_a$$

$$s^2_b \rightarrow \sigma^2_e + \sigma^2_{aDG} + C\sigma^2_{aD} + \frac{RC}{R-1} \sum D^2_b$$

$$s^2_c \rightarrow \sigma^2_e + \sigma^2_{aDG} + B\sigma^2_{aG} + \frac{RB}{C-1} \sum G^2_c$$

$$s^2_{rb} \rightarrow \sigma^2_e + \sigma^2_{aDG} + C\sigma^2_{aD}$$

$$s^2_{rc} \rightarrow \sigma^2_e + \sigma^2_{aDG} + B\sigma^2_{aG}$$

$$s^2_{bc} \rightarrow \sigma^2_e + \sigma^2_{aDG} + \frac{R}{(B-1)(C-1)} \sum \sum (DG)^2_{bc}$$

$$s^2_{rbc} \rightarrow \sigma^2_e + \sigma^2_{aDG}$$

Scrutiny of the foregoing expected values indicates that  $s^2_{rbc}$  is appropriate for testing all three 2-way interactions, that  $s^2_c$  should be tested against  $s^2_{rc}$ , and  $s^2_b$  against  $s^2_{rb}$ . No test of  $s^2_r$  is possible, but this is not serious since it would only be a test of the significance of individual differences.

If we had replication of measurements (each person measured  $m$  times under each of the  $BC$  conditions), we would have an  $s^2_{re}$ , as an estimate of  $\sigma^2_e$ , which would permit a test of the 3-way interaction effect. If  $F_{rbc}$  were significant it would only mean that 3-way interaction is, for our sample of  $R$  persons, greater than expected on the basis of errors of measurement. As usual (when an estimate of  $\sigma^2_e$  is used as the error term), no generalization to a population of individuals is possible. It is a fact, however, that 3-way interactions involving individuals are usually significant; hence having measurement replication does not change the procedure, for choosing the error term, from that depicted in the just previous paragraph.

**Case XV.** Mixed model [ $adG$ ], with 1 score per cell. Situations calling for this model in psychology are not plentiful. Suppose  $R$  children are observed under  $C$  different social situations by  $B$  observers, each of whom rates (on a 10-point scale) each child for a particular aspect of behavior, e.g., social participation. Primary interest would be in the effect of conditions (the  $G_c$  effects) with secondary interest in observer bias (the raters being regarded as a sample of observers having  $d_b$  "effects") and possible interest in 2-way interaction effects. For model [ $adG$ ] the meaning of the several variance estimates is, aside from a common  $\sigma^2_e$  term, as follows:

$$s^2_r \rightarrow \sigma^2_{adG} + C\sigma^2_{ad} + B\sigma^2_{aG} + BC\sigma^2_a$$

$$s^2_b \rightarrow \sigma^2_{adG} + C\sigma^2_{ad} + R\sigma^2_{dG} + RC\sigma^2_d$$

$$s^2_c \rightarrow \sigma^2_{adG} + B\sigma^2_{aG} + R\sigma^2_{dG} + \frac{RB}{C-1} \sum G^2_c$$

$$s^2_{rb} \rightarrow \sigma^2_{adG} + C\sigma^2_{ad}$$

$$s^2_{rc} \rightarrow \sigma^2_{adG} + B\sigma^2_{aG}$$

$$s^2_{bc} \rightarrow \sigma^2_{adG} + R\sigma^2_{dG}$$

$$s^2_{rbc} \rightarrow \sigma^2_{adG}$$

Obviously, the appropriate error term for testing all three 2-way interactions is  $s^2_{rbc}$ , but trouble is encountered in finding an error term for testing the main effects. Keeping in mind that a test of  $s^2_r$  is of trivial importance, we note that if the  $(dG)_{bc}$  interaction effect were zero,  $s^2_b$  could be tested against  $s^2_{rb}$  and  $s^2_c$  against  $s^2_{rc}$ ; or if the  $(ad)_{rb}$  interaction were zero,  $s^2_{bc}$  could be used to test  $s^2_b$ ; and if interaction  $(aG)_{rc}$  were zero,  $s^2_c$  could be tested by using  $s^2_{bc}$ . One can scarcely make a priori the assumption that any of these 2-way interactions is zero; in fact, the safest presumption is that none of the 3 is zero. It is frequently asserted that the failure of a 2-way interaction to be significant when tested against  $s^2_{rbc}$  can be used to justify the assumption of a zero interaction, but failure to be significant means only that it could be zero. Furthermore, if  $R$  and  $B$  are small an interaction would need be sizable to be detected. This issue, along with a similar one, will be discussed later under the heading "Preliminary tests." Suffice it to say now that model  $[adG]$  is not recommended.

**Case XVI.** Components of variance model  $[adg]$ . If anyone finds a situation in which all 3 bases of classification involve sampling, he will need to know that for model  $[adg]$  the variance estimates have expected values specifiable from those of model  $[adG]$  by replacing  $G$  with  $g$  except for the fourth term of  $s^2_c$ , which becomes  $RB\hat{\sigma}_g^2$ . The 2-way interactions can be tested against  $s^2_{rbc}$ , but there is no way of testing the main effects without making precisely the same assumptions regarding 2-way interactions that were indicated for Case XV.

**Case XVII.** Mixed model  $[aDG]$ , but a pseudo 3-way classification. Suppose a sample of  $R$  individuals in block 1, a sample of  $R$  different individuals in block 2, and so on. The  $B$  blocks represent  $B$  experimental conditions, the effects of which are to be determined, and at the same time the  $C$  columns stand for another factor which is also to be evaluated. The  $B$  sets of  $R$  individuals are used because it is not feasible to use each person under each block condition. Or suppose the blocks stand for different groups (say, diagnostic) from each of which  $R$  cases are drawn at random. We wish to compare the groups and also the  $C$  conditions.

Let us re-examine Table 51 in order to determine how to set up the model for this situation. We first note that for Case XIV the variation among the row means ( $\bar{X}_{r..}$ ) contributes to  $s^2_r$  as an estimate of individual difference variation, whereas for Case XVII

each of these row means is an average for  $B$  different individuals; hence row means do not hold for individuals. We do, however, have individual difference variation within each block, as represented by means of the type  $\bar{X}_{r \cdot b}$ . (right-hand part of Table 51). Accordingly, we can anticipate a sum of squares for individual differences which will involve combining the sum of squares within each block; i.e.,  $C \sum \sum (\bar{X}_{r \cdot b} - \bar{X}_{\cdot \cdot b})^2$ , with  $RB - B$  degrees of freedom. The resulting variance estimate may be labeled  $s_i^2$ , for individual differences.

In ordinary 3-way classification (Case XIV) the  $B$  sets of means of the type  $\bar{X}_{r \cdot b}$  have to do with row (individual) by block interaction, an interaction which reflects the failure of the individuals to maintain similar score positions from block to block. But with independent cases in each block, no block by row interaction is possible; a person can't react differently from 1 block to another unless he has been measured under more than 1 block condition. Consider next the  $\bar{X}_{r \cdot c}$  type of mean at the bottom of Table 51. These means ordinarily enter into row by column interaction, but in the present case each of these means is the average for  $B$  different individuals who just happened to have been assigned the same row number. Therefore, there can be no row by column interaction in the usual sense. We have, nevertheless,  $RB$  independent individuals in a total of  $RB$  (instead of  $R$ ) rows; hence there could be a meaningful individual by column interactive effect (not testable with 1 score per cell, but present as a source of variation).

What of a possible 3-way interaction involving rows, blocks, and columns? This does not make sense since an individual can in no way react inconsistently from 1 block condition to another without having been subjected to different block conditions.

With the foregoing in mind, we may write the following specific model for Case XVII:

$$(X_{rbc} - \hat{\bar{X}}) = a_i + D_b + G_c + (DG)_{bc} + h_{rbc}$$

in which  $a_i$  indicates individual difference effects and  $h_{rbc}$  is the remainder after the first 4 parts have been subtracted from  $(X_{rbc} - \hat{\bar{X}})$ . The several sums of squares and their  $df$ 's are given in Table 57. Note how the first line differs from the first line of Table 52; note also the similarity of the remainder sum of squares to the remainder (or last) term in equation (101), p. 286, and to

Table 57. MODIFICATION OF VARIANCE TABLE 52 FOR CASE XVII:  
*R* DIFFERENT AND INDEPENDENT INDIVIDUALS IN EACH BLOCK

Source	Sum of Squares	<i>df</i>	Variance Estimate
Individuals *	$\sum_b \sum_c R \sum_i (\bar{X}_{rb.} - \bar{X}_{..b})^2$	$RB - B$	$s^2_i$
Blocks	$RC \sum_b (\bar{X}_{..b} - \bar{X})^2$	$B - 1$	$s^2_b$
Columns	$RB \sum_c (\bar{X}_{.c.} - \bar{X})^2$	$C - 1$	$s^2_c$
<i>B</i> × <i>C</i> interaction	$R \sum_b \sum_c (\bar{X}_{.bc} - \bar{X}_{..b} - \bar{X}_{.c.} + \bar{X})^2$	$(B - 1)(C - 1)$	$s^2_{bc}$
Remainder	$\sum_b \sum_c R \sum_i (X_{rbci} - \bar{X}_{rb.} - \bar{X}_{.bc} + \bar{X}_{..b})^2$	$B(R - 1)(C - 1)$	$s^2_h$
Total	$\sum_b \sum_c R \sum_i (X_{rbci} - \bar{X})^2$	$RBC - 1$	

\* The sum of squares for individuals is computed by substituting in  $\frac{1}{RC} [R \sum_b \sum_c (\sum_i X_{rbci})^2 - \sum_b \sum_c R (\sum_i X_{rbci})^2]$ .

the row by column interaction term in Table 46. Actually, the remainder in Table 57 involves possible individual by column interaction, composed of ordinary row by column interaction within each block, then summed over blocks.

The expected values of the several variance estimates are as follows (recall that  $s^2_i$  contains  $\sigma^2_e$  as a component):

$$s^2_i \rightarrow \sigma^2_i + \sigma^2_{aG}$$

$$s^2_b \rightarrow \sigma^2_i + \sigma^2_{aG} + \frac{RC}{B - 1} \sum_b D^2_b$$

$$s^2_c \rightarrow \sigma^2_e + \sigma^2_{aG} + \frac{RB}{C - 1} \sum_c G^2_c$$

$$s^2_{bc} \rightarrow \sigma^2_e + \sigma^2_{aG} + \frac{R}{(B - 1)(C - 1)} \sum_b \sum_c (DG)^2_{bc}$$

$$s^2_h \rightarrow \sigma^2_e + \sigma^2_{aG}$$

From these values we see at a glance that  $s^2_i$  is the error term for testing  $s^2_b$ , a test which is analogous to  $s^2_b/s^2_w$  in the 1-way classification setup for the difference between the means of in-

dependent groups. For testing  $s^2_c$  the remainder estimate,  $s^2_h$ , is appropriate. Since  $s^2_h$  is, in part, an estimate of individual by column interaction, we find an analogue in the 2-factor setup (Case VII, p. 309) for which a row by column interaction provides the correct variance estimate for testing column effects when the column means are correlated (based on the same individuals).

The remainder variance estimate is also appropriate for testing the  $B \times C$  interaction. This interaction has a special meaning when  $B$  stands for different groups and  $C$  stands for  $C$  tests all scored in comparable standard score form. The column means for each block are the basis for a given group's profile; hence a test of the  $B \times C$  interaction tells us whether there are significant differences among the profiles for the  $B$  groups.

Caution: Case XVII as here outlined calls for the same number of individuals per block (or group).

**Assumption of homogeneity of variance.** Cases X, XI, and XII require similar individual difference variance for all cubicles, but only Case X permits a test of the assumption. For Cases XIII, XIV, XV, and XVI it is assumed that error of measurement variance is the same from cubicle to cubicle. The assumption for these cases is not testable unless one has measurement replication with  $m$  scores per cubicle. Case XVII assumes that the row variance within blocks is homogeneous from block to block when using  $s^2_i$  to test  $s^2_b$ , and that the row by column interaction within blocks is similar from block to block when testing either  $s^2_c$  or  $s^2_{bc}$  against  $s^2_h$ . Both of these assumptions are testable since the required within-block estimates can be computed.

## PRELIMINARY TESTS AND POOLING

When we discussed Case XV, we found that certain effects could not be tested without assuming that an interaction is zero. The temptation is to assume an interaction is zero if it fails to be significant when tested against an appropriate error term. The writers of textbooks on mathematical statistics are remarkably mum on this point, presumably because the situation gets too "iffy": a main effect is significant *if* it reaches, say, the .05 level, and *if* a certain interaction was not significant at a specified level. Under such circumstances a  $P$  for an effect ceases to have the same meaning as when unencumbered by conditional probabilities.



Note that preliminary tests may have to do with the assumption of zero interaction in the numerator term of  $F$  (as for Case XV) or in the denominator term (as for Cases II and XI). Failure to satisfy the assumption of a zero interaction in the numerator will lead to too many "significant"  $F$ 's. Stated differently, significance for a main effect cannot be safely claimed because the numerator involves a possible confounding of interactive and main effects. As pointed out earlier, failure to satisfy an assumption of zero interaction in the denominator term will lead to too few significant  $F$ 's, which means that an obtained  $F$  possesses greater significance than its  $P$  indicates.

Preliminary tests are also used in connection with the "pooling" of sums of squares and of their  $df$ 's. To understand the meaning of pooling, let us consider Case X in which all effects are testable against  $s^2_{...}$ . The advocated steps are: First,  $s^2_{abc}$  is tested against  $s^2_{...}$ . If this  $F$  is not significant at, say, the .05 level, the sum of squares for the 3-way interaction term is combined with that of  $s^2_{...}$ , with the  $df$ 's also being summed. Dividing the pooled sum by the pooled  $df$  gives another estimate of variance for the error term. This estimate is next used to test the 2-way interactions, which if insignificant provide additional sums of squares and  $df$ 's for adding to the pool already made up.

The claimed advantage of pooling is that the number of degrees of freedom for the denominator, or error, term of  $F$  is thereby increased, with a resultant more stable estimate of variance. But whether this procedure provides an improved or better estimate depends, of course, on whether the interactions judged to be insignificant are really zero in the sampled population. Actually, the  $F$  based on the pooled values may be either larger or smaller than the  $F$  based on the appropriate variance estimate obtained without pooling. When one examines the  $F$  table, one sees that the gain in  $df$  does not have an appreciable effect, in the sense that a smaller  $F$  is required for significance, except when  $n_2$  is very small, say less than 8 or 10. It should be clearly noted that the gain in  $df$  by pooling does not lead to a reduction in the sampling errors of the means being tested.

The use of preliminary tests as a basis for pooling is not nearly so defensible as textbooks written prior to 1951 would have us believe. The work of Paull ‡ indicates that the usually advocated

‡ Paull, A. E., On a preliminary test for pooling mean squares in the analysis of variance, *Annals math. Stat.*, 1950, **21**, 539-556.

rule (that when  $F$  is less than the value required for the .05 level, pooling is permissible and advisable) is far from satisfactory. He sets up an elaborate set of rules leading to the decision "never pool" or "sometimes pool" or "always pool." Space does not permit an exposition of his rules here. A simple rule to follow when the  $df$ 's are equal, or when unequal provided both are greater than 6, is to pool only when  $F$  is less than 2. Even when one follows the rules,  $F$ 's based on pooling do not lead to  $P$ 's of precisely the same meaning as  $P$ 's obtained from  $F$ 's which do not involve pooling.

### HIGHER-ORDER CLASSIFICATION

There are times when it is both desirable and feasible to study the variations of a dependent variable associated with variations in more than 3 variables. For such a study the data are classifiable in more than 3 ways. We have already mentioned the setup in which an observation is made on each of  $m$  individuals under each of the combinations of conditions defined by rows, blocks, and columns. There will be  $RBC$  scores for each individual, and the scores may be classified not only as belonging to a given row and a specified column of a particular block but also as belonging to a certain individual. Although it is easy to make an orderly arrangement of the data for quadruple classification, the required computations become somewhat burdensome. For the situation involving a fourth classification, based on either individuals or on a fourth independent variable, there will be 16 sums of squares: 1 for total, 4 for between groups, 6 for simple interactions, 4 for triple interactions, and 1 for quadruple interaction. When 5 classifications are used we will have sums of squares for the total, 5 betweens, 10 simple interactions, 10 triple interactions, 5 quadruple interactions, and 1 fifth-order interaction. It is not within the scope of this book to outline the computations for these higher-order classifications. §

The possibilities of the variance technique as a method of extracting from 1 set of data information regarding not only primary effects but also interactions have, at times, led to rather indiscriminate inclusions of variables. For instance, a classification of subjects as male or female may be made in order to determine possible sex differences. Since the typical experiment for

§ See Edwards, A. L., and Horst, P., The calculation of sums of squares for interaction in the analysis of variance, *Psychometrika*, 1950, 15, 17-24.

which the variance technique is used is likely to be based on a relatively small number of subjects, it is very doubtful whether any information of value will be added to the sum total of the already inconsistent findings concerning sex differences.

Those who carry out studies involving more than triple classification encounter great difficulty in interpreting significant higher-order interactions. Some have thought it safe, after ascertaining the sums of squares for the primaries and the simple and triple interactions, to use the remainder variance, which is a composite of untested higher-order interactions, as an error term. Such a practice assumes insignificance for the interactions whose sums of squares are thus allowed to combine, but since there are instances of significant quadruple interaction, the cautious investigator will extract and test all the possible interactions before using such a remainder as the error term for  $F$ .

As a matter of fact, the choice of the proper error term for higher-order classifications is, at times, quite complicated. For the simple 4-way setup involving the fixed constants model  $[ADGH]$ , with  $m$  replications of individuals per cubicle, the  $s^2_w$  estimate is the correct error term for testing all 4 main effects and all 11 interactions. For the mixed model  $[aDGH]$ , with  $a_r$  standing for individuals (a typical setup), the main effects are tested against the respective 2-way interactions involving individuals (as in Case XIV, p. 330); the 3-way interactions are tested against the 4-way interaction, but there are no exact tests of the three 2-way interactions involving individuals (which are usually significant when testable). For further detail on the 4-way classification, the reader is referred to the Mentzer paper cited on p. 304.

### FACTORIAL AND LATIN SQUARE DESIGNS

The student who encounters the term "factorial design" will need to know that it is difficult to make a distinction between factorial design and the analysis of variance setups discussed in this chapter. The bases for classification are referred to as factors; the categories within a classification are termed "levels." Perhaps the term factorial design is inappropriate when 1 basis for classification is persons.

The Latin square design had its origins in agricultural experimentation. If  $T$  different treatments (fertilizers) are to be evalu-

ated, a plot of land is laid off into  $T$  rows and  $T$  columns and the treatments are so assigned that each treatment occurs only once in each row and only once in each column. With Latin letters standing for the treatments, one might have the accompanying square, an examination of which reveals that this is a scheme for

		Columns			
		1	2	3	4
Rows	I	A	D	B	C
	II	B	A	C	D
	III	C	B	D	A
	IV	D	C	A	B

balancing out the effects of possible fertility differentials from row to row and also from column to column.

Some researchers in psychology have used the Latin square principle as a way of balancing the effect of individual differences and order of testing (practice). That is, with  $T$  conditions to be evaluated, the rows stand for  $T$  individuals and the columns for  $T$  orders of testing, with Latin letters representing the  $T$  conditions. The design also can be and has been used in lieu of a complete 3-way factorial design when all 3 factors involve the same number of levels. For example, 16 properly arranged observations may be used instead of the 64 observations required for a complete 3-way classification plan with 4 levels per classification. This second use of the Latin square principle is not for the purpose of balancing out the effect of a factor but rather for evaluating the effect of factors which are deliberately varied.

Thus, it would seem that the Latin square design might be very useful in psychology, but before we accept it uncritically (as some advocates have), we need to examine the underlying mathematical model, which may be written as

$$(X_{rel} - \hat{X}) = \alpha_r + \delta_c + \gamma_t + f_{rel}$$

The  $\alpha$ ,  $\delta$ , and  $\gamma$  refer to row, column, and treatment effects, and  $f_{rel}$  is a remainder, or residual. It follows from the model that the breakdown of the total sum of squares and degrees of freedom will lead to sums of squares for rows, for columns, and for treatments, each with  $T - 1$  degrees of freedom. These sums of squares will use up  $3T - 3$  of the total  $df$ ,  $T^2 - 1$ ; hence there

remain  $T^2 - 3T + 2$  degrees of freedom for the residual sum of squares. The variance estimate based on the residual is used as the error term (denominator) of  $F$  when testing  $s^2_r$ ,  $s^2_c$ , and  $s^2_t$ .

When the foregoing model is compared with that of complete 3-way classification (p. 327), we see a marked difference: the absence of interaction terms. *For the Latin square design it is assumed that all interactions are zero.* This assumption is necessary because there are not enough degrees of freedom available for taking out possible interactive effects in order to arrive at an error variance estimate appropriate for  $F$ . The more important thing here is not that the Latin square does not permit a test of the interactions but that this design does not provide a proper error term unless the interactions are in fact zero.

Why doesn't the residual provide a suitable error term when 1 or more of the interactions is not zero? In considering this question we must distinguish between the fixed constants model [ADG], which is applicable when the Latin square is used in place of a complete 3-way factorial design, and the mixed model [aDG], which is called for when, say, rows stand for individuals.

We have already seen that for the mixed model the 2 main effects are tested by  $F_c = s^2_c/s^2_{rc}$  and  $F_b = s^2_b/s^2_{rb}$  when we have a complete 3-way classification (Case XIV, p. 330). If for Case XIV we pooled the sums of squares for the three 2-way interactions and the 3-way interaction we would have a residual term exactly analogous to the residual of the Latin square design. Note that the variance estimate obtained by pooling sums of squares and  $df$ 's (Case XIV) is equivalent to calculating a weighted average of the 4 interaction variance estimates. That is, the pooled sum of squares is equal to

$$(R-1)(B-1)s^2_{rb} + (R-1)(C-1)s^2_{rc} \\ + (B-1)(C-1)s^2_{bc} + (R-1)(B-1)(C-1)s^2_{rbc}$$

which, when divided by the sum of the degrees of freedom (sum of the weights), gives the residual variance estimate.

Will such an estimate, as the error term, be larger or smaller than the estimates,  $s^2_{rb}$  and  $s^2_{rc}$ , which are the proper error terms for testing  $s^2_b$  and  $s^2_c$ ? If there are 2-way interactive effects present, the value of  $s^2_{rbc}$  will tend in general to be smaller than either  $s^2_{rb}$  or  $s^2_{rc}$ . Accordingly, when we pool, or take a weighted



average, we will have an average which has been pulled down by the relatively small value of  $s^2_{rbc}$ . (It would also be affected by the presence or absence of  $B \times C$  interaction.) This means that the variance estimate based on the pooled residual will tend to be smaller than either  $s^2_{rb}$  or  $s^2_{rc}$ ; hence the use of the residual variance as *the* error term for Case XIV will produce too many "significant"  $F$ 's. Likewise, and for precisely the same reasons, the use of the residual variance of the Latin square design as the error term for testing main effects will produce too many "significant"  $F$ 's if the 2-way interactions involving individuals are nonzero. This is true because, in effect, the residual variance of the Latin square is a weighted average analogous to that obtained by pooling for Case XIV. When it is recalled that tested interactions involving individuals are nearly always significant, we see that  $F$ 's derived from Latin squares (mixed model) are not only not dependable but also apt to be fallacious.

Let us next consider Latin squares used in lieu of 3-way factorial designs when all the effects involve the fixed constants model [ADG]. It will be recalled that for complete 3-way classification the proper error variance for testing all effects is  $s^2_w$  when there are  $m$  cases per cubicle (Case X, p. 328), and  $s^2_{rbc}$  when we do not have replication, provided it can be assumed that the 3-way interaction is zero (Case XI). If for the 3-way factorial setup we extracted the sum of squares for each of the 3 main effects, and then took the remainder as a residual we would, of course, have the exact equivalent of a pooled sum of squares involving the three 2-way interactions, the 3-way interaction, and the within-cubicles variation (if we had replication). As in the mixed model, the variance estimate based on such a residual will be a weighted average of  $s^2_{rb}$ ,  $s^2_{rc}$ ,  $s^2_{bc}$ ,  $s^2_{rbc}$ , and  $s^2_w$ . But this time the weighted average will tend to be higher than the proper error variance,  $s^2_w$  if any 1 of the 4 interactions is not zero in the population. Thus, for the fixed constants model the use of the residual variance estimate as the denominator of  $F$  will tend to give too few "significant"  $F$ 's. This tendency to underestimate significance will become greater as more of the 4 interactions fail to be zero, and furthermore the larger the interaction(s) the greater will be the underestimation.

Since the residual variance in the Latin square is precisely analogous to the weighted average just discussed, we have the



unescapable conclusion that too few significant  $F$ 's will emerge when the Latin square design is used in place of a 3-way factorial design (fixed constants model) if the assumption of zero interaction does not hold for any of the 4 possible interactions. Since the Latin square design does not provide data for testing the assumption of zero interactions, its use cannot be defended unless there are strong a priori reasons for believing that *all* 4 interactions are really zero. The only consolation left for the user of the Latin square (fixed constants model) is that an obtained significant  $F$  may possess greater significance than indicated by the  $F$  table, but even this solace is ephemeral if it occurs to him that the assumptions might just happened to have been met. Actually, though one can trust a significant  $F$ , one cannot safely claim added significance unless there is reason for suspecting real interaction. The most telling objection to the fixed constants Latin square is that its use stacks the cards against the obtaining of significant  $F$ 's—the null hypothesis will all too often be falsely accepted.

## Analysis of Variance: Covariance Method

It is usually possible in experimentation to choose, either by random methods or by pairing or matching, groups that are comparable on variables judged relevant to the comparisons to be made. There are times, however, when it is more practicable to use intact groups which may differ in important respects, and occasionally one may wish to make an unanticipated comparison which does not seem justifiable in light of known differences between groups. Experimental control is the ideal, but, if this cannot be attained, one may resort to statistical allowances and thereby arrive at valid conclusions.

Suppose that 2 intact groups are being used to evaluate the relative merits of 2 methods of memorizing and that the mean IQ is 105 for group A and 111 for group B. Now, if there is an appreciable correlation between the particular memorizing ability involved and intelligence, the results will need qualifying because of the difference in intelligence of the 2 groups. It would seem logical to use the regression equation, for estimating memory score from intelligence, as a basis for predicting how much of a difference in memorizing would arise because of the group difference in IQ's. Let us suppose that the mean memory performance is 60 for group A and 70 for group B, and that substituting 105 and 111 in the regression equation yields a predicted value of 62 for group A and of 68 for group B. Thus our prediction would lead us to expect a difference of 6 points, and accordingly it would be said that 6 of the obtained difference of 10 could be attributed to lack of comparability of the 2 groups with respect to intelligence.

The next question concerns the proper sampling error to use in evaluating the adjusted difference. It should be obvious that the ordinary procedure is inapplicable for the simple reason that

we have tampered with the obtained means and in so doing have interfered somewhat with the operation of chance.

It is the purpose of this chapter to give a precise method for making allowance for an uncontrolled variable and to set forth the sampling error adjustment which is needed in testing the statistical significance of the difference between "corrected" means. The method is applicable whenever it seems desirable to correct a difference on a dependent variable for a known difference on another variable which for some reason could not be controlled by matching or by random sampling procedures. Since the scheme about to be proposed has an analysis of variance setting, the reader can readily guess that it will provide an adjustment for, and a test of significance of, the differences between two or more groups, and that it will be usable for either large or small samples. It is assumed that the dependent variable has a distribution which does not depart too far from the normal type and that the variances from group to group are similar.

In order to present the required adjustments, we need first to consider *covariance*, which is defined as  $\Sigma xy/N$  or  $\Sigma(X - \bar{X})(Y - \bar{Y})/N$ . The sum of products of deviations can be broken down into components in a manner similar to that used with a sum of squares. In the simplest situation we can have  $m$  pairs of  $X$  and  $Y$  scores in each of  $k$  groups. These pairs of scores can be recorded in some such fashion as that depicted in Table 58. Note that  $X_{ij}$  and  $Y_{ij}$  stand for the  $X$  and  $Y$  values

Table 58. SCHEMA OF SCORES FOR COVARIANCE

Group							
1		2		$j$		$k$	
$X_{11}$	$Y_{11}$	$X_{12}$	$Y_{12}$	$X_{1j}$	$Y_{1j}$	$X_{1k}$	$Y_{1k}$
$X_{21}$	$Y_{21}$	$X_{22}$	$Y_{22}$	$X_{2j}$	$Y_{2j}$	$X_{2k}$	$Y_{2k}$
$X_{i1}$	$Y_{i1}$	$X_{i2}$	$Y_{i2}$	$X_{ij}$	$Y_{ij}$	$X_{ik}$	$Y_{ik}$
$X_{m1}$	$Y_{m1}$	$X_{m2}$	$Y_{m2}$	$X_{mj}$	$Y_{mj}$	$X_{mk}$	$Y_{mk}$

of the  $i$ th individual in the  $j$ th group. Note also that in allowing  $i$  to take on values running from 1 to  $m$  we do not imply any order for the individual, and that the  $i$ th individual in one group is in no sense paired with the  $i$ th case in another group. The product of the deviation scores for the  $i$ th individual in the  $j$ th group would be  $(X_{ij} - \bar{X})(Y_{ij} - \bar{Y})$ , in which  $\bar{X}$  and  $\bar{Y}$  are the

means for all  $km$  cases. The total sum of products would be  $\sum_{i,j} (X_{ij} - \bar{X})(Y_{ij} - \bar{Y})$ . Now each deviation can be expressed in terms of two components in exactly the same way as in Chapter 15; i.e., one part is the deviation of the score from the mean of the group to which it belongs, and the other part is the deviation of the group mean from the total mean. Thus we have

$$(X_{ij} - \bar{X}) = (X_{ij} - \bar{X}_j) + (\bar{X}_j - \bar{X})$$

and

$$(Y_{ij} - \bar{Y}) = (Y_{ij} - \bar{Y}_j) + (\bar{Y}_j - \bar{Y})$$

Then the above sum of the products becomes

$$\sum_{i,j} [(X_{ij} - \bar{X}_j) + (\bar{X}_j - \bar{X})][(Y_{ij} - \bar{Y}_j) + (\bar{Y}_j - \bar{Y})]$$

When the bracketed expressions are multiplied together, four terms result, and, since two of these vanish, we have left that the total sum of products is equal to

$$\sum_{i,j} (X_{ij} - \bar{X}_j)(Y_{ij} - \bar{Y}_j) + m \sum_j (\bar{X}_j - \bar{X})(\bar{Y}_j - \bar{Y})$$

The first of these terms involves a *within*-groups sum of products, whereas the second is for *between* groups. If there happens to be an unequal number of cases per group, the  $m$  of the second term goes under the summation sign as  $m_j$ . The degrees of freedom for the total sum of products is  $km - 1$ , or  $N - 1$ , where  $N$  is the sum of the  $m_j$ 's; the  $df$ 's for the within and between terms are  $km - k$  (or  $N - k$ ) and  $k - 1$  respectively.

It will be of convenience to assemble in a table the sums of products, along with the sums of squares, for both the  $X$  and  $Y$  variables. These will be found in the first three lines of Table 59.

Although we are here presenting the covariance technique as a method for making such adjustments as discussed in introducing this chapter, it is of interest to link covariance with the problem of correlation. The product moment correlation coefficient is usually defined as

$$r = \frac{\sum xy}{N\sigma_x\sigma_y}$$

which may be written as

$$r = \frac{\sum xy}{N\sqrt{\frac{\sum x^2}{N}}\sqrt{\frac{\sum y^2}{N}}} = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2}\sqrt{\sum(Y - \bar{Y})^2}}$$

Table 59. SETUP FOR ANALYSIS OF VARIANCE BY COVARIANCE ADJUSTMENTS

	Total	Within	Between
1. Sum of products	$\sum_i \sum_j (X_{ij} - \bar{X})(Y_{ij} - \bar{Y})$ (A <sub>i</sub> )	$\sum_i \sum_j (X_{ij} - \bar{X})(Y_{ij} - \bar{Y})$ (A <sub>w</sub> )	$\sum_i m_j (\bar{X}_j - \bar{X})(\bar{Y}_j - \bar{Y})$ (A <sub>b</sub> )
2. Sum of squares for X's	$\sum_i \sum_j (X_{ij} - \bar{X})^2$ (B <sub>i</sub> )	$\sum_i \sum_j (X_{ij} - \bar{X})^2$ (B <sub>w</sub> )	$\sum_i m_j (\bar{X}_j - \bar{X})^2$ (B <sub>b</sub> )
3. Sum of squares for Y's	$\sum_i \sum_j (Y_{ij} - \bar{Y})^2$ (C <sub>i</sub> )	$\sum_i \sum_j (Y_{ij} - \bar{Y})^2$ (C <sub>w</sub> )	$\sum_i m_j (\bar{Y}_j - \bar{Y})^2$ (C <sub>b</sub> )
4. df	N - 1	N - k	k - 1
5. Correlation coefficient	$\frac{A_t}{\sqrt{B_t} \sqrt{C_t}}$	$\frac{A_w}{\sqrt{B_w} \sqrt{C_w}}$	$\frac{A_b}{\sqrt{B_b} \sqrt{C_b}}$
5a. df for r	N - 2	N - k - 1	k - 2
6. $b_{xy}$	$A_t/C_t$	$A_w/C_w$	$[A_b/C_b]$
7. Adjusted $\Sigma x^2$	$(B_t - A_t^2/C_t)$	$(B_w - A_w^2/C_w)$	adjusted $B_b$
8. df	N - 2	N - k - 1	k - 1

or as a function of a sum of products and two sums of squares. Using the sums of Table 59, we may specify three correlations: one based on the total sums, one based on the within sums, and one based on the between sums. These three correlations are indicated in line 5 by letters **A**, **B**, and **C**, with appropriate subscripts used to designate the several sums in the first three lines of the table. Line 5a gives the  $df$ 's for the  $r$ 's.

Note that the between-groups  $r$  is actually the correlation between the  $X$  means and the  $Y$  means for the groups. If this  $r$  is significant, it follows that one source of the correlation for the total group is the heterogeneity resulting from the throwing together of groups with unlike means. (This between-groups correlation is meaningless when only two groups are involved. Why?) Stated differently, an appreciable between-groups  $r$  indicates that the total  $r$  is spurious; this spuriousness is eliminated when  $r$  is computed from the within sums. The similarity of the within-groups  $r$  to the partial correlation coefficient will be recognized by the discerning student, especially if he recalls the derivation of the latter.

We now turn to the use of covariance as a basis for allowing for the influence of an uncontrolled variable on the differences between group means. The question here is not what the result would be if the uncontrolled variable were held constant, as in partial correlation, but rather what the result would be if the groups were made comparable with respect to the uncontrolled variable. Let  $X$  represent the dependent variable, and  $Y$  the uncontrolled variable. It is presumed that the  $\bar{Y}_j$  values differ, and that  $X$  is correlated with  $Y$  in a linear fashion. For purposes of exposition we shall refer to Table 59, which will serve as an outline of the required computations. Line 6 of this table gives the regression coefficients ( $b_{xy}$ ) for predicting  $X$  from  $Y$ . Since no use will be made of  $A_b/C_b$ , it is bracketed; it need not be computed.

That these  $A/C$  values are regression coefficients can readily be demonstrated. In Chapter 7 the regression of  $X$  on  $Y$  was given as

$$b_{xy} = r \frac{\sigma_x}{\sigma_y}$$



Since, as we have seen above,

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2} \sqrt{\Sigma y^2}}, \quad \sigma_x = \sqrt{\Sigma x^2 / N}, \quad \text{and} \quad \sigma_y = \sqrt{\Sigma y^2 / N}$$

we have

$$b_{xy} = \frac{\Sigma xy}{\sqrt{\Sigma x^2} \sqrt{\Sigma y^2}} \cdot \frac{\sqrt{\Sigma x^2 / N}}{\sqrt{\Sigma y^2 / N}}$$

$$= \frac{\Sigma xy}{\Sigma y^2} = \frac{A}{C}$$

In order to make allowance for the uncontrolled differences in  $\bar{Y}_j$ , we need not only to adjust the  $\bar{X}_j$  values but also to make an adjustment to the error term, which is used as the denominator of the  $F$  ratio in testing the difference between the adjusted  $X$  means. As in the simpler situation of Chapter 15,  $F$  will involve the ratio of a between-groups to a within-groups variance estimate.

First, let us consider the method of making the adjustment to the total and to the within-groups variance estimates. The problem here is that of specifying how much of the variation in  $X$  can be predicted from variation in  $Y$  and then of subtracting this to secure the left-over variation as an adjusted value. But this left-over variance is nothing more than the residual variance, or square of the standard error of estimate, obtainable from formula (35):

$$\sigma^2_{x \cdot y} = \sigma^2_x - r^2 \sigma^2_x$$

Actually the adjustment is to be made to the sum of squares. In order to state the residual variance in terms of sums, we may substitute for  $\sigma^2_x$  and  $r^2$ . Thus,

$$\sigma^2_{x \cdot y} = \frac{\Sigma x^2}{N} - \frac{(\Sigma xy)^2}{(\Sigma x^2)(\Sigma y^2)} \cdot \frac{\Sigma x^2}{N}$$

hence,

$$N\sigma^2_{x \cdot y} = \Sigma x^2 - \frac{(\Sigma xy)^2}{\Sigma y^2}$$

Since  $N\sigma^2$  always equals a sum of squares, the value of  $N\sigma^2_{x \cdot y}$  is obviously the sum of squares for the residuals. In the notation

of this chapter,

$$N\sigma^2_{x \cdot y} = \sum \sum (X_{ij} - \bar{X})^2 - \frac{[\sum \sum (X_{ij} - \bar{X})(Y_{ij} - \bar{Y})]^2}{\sum \sum (Y_{ij} - \bar{Y})^2}$$

would be the residual sum of squares after the regression adjustment. This sum can be written as

$$N\sigma^2_{x \cdot y} = B_t - \frac{A_t^2}{C_t}$$

which is the entry for the *total* group in line 7 of Table 59. Similarly, the corresponding residual, or adjusted, sum of squares for *within* groups is  $B_w - A_w^2/C_w$ .

At first thought it would seem logical to adjust  $B_t$  by the use of  $A_t$  and  $C_t$ , but the between-groups correlation (and regression) is affected by the differences between the  $X$  means, which are the differences to be adjusted and then tested for statistical significance. Our adjustment should be one which is independent of the differences to be tested. This suggests that the regression for within groups, or  $A_w/C_w$ , should be used since the regression for the total is also affected by the difference which we are out to test. In so far as we are concerned solely with the adjustment of the between-groups  $X$  means, the best adjustment would be by means of the within-groups regression. This could take the form of either an adjustment to the between-groups sum of squares for  $X$  or a direct adjustment to the several  $\bar{X}_j$  values.

Although the latter would be the best way of ascertaining how much of an effect the noncomparability of the groups with respect to  $Y$  had upon the  $X$  means, there is another consideration as to whether the within regression is appropriate for adjusting the between-groups sum of squares. It will be recalled that  $F$  is to be taken as the ratio of a variance estimate based on the between sum of squares to that based on within groups, and that the two variance estimates being so compared must be independent estimates. Now, if we adjust both the within and the between sum of squares by means of the same regression coefficient (say, that based on within groups), any sampling error in this regression coefficient would have a similar effect on both adjustments; hence it could not be argued that the resulting adjusted sums of squares

possess the requisite independence. Therefore variance estimates based thereon would not be strictly independent.

This difficulty is overcome by taking the adjusted sum of squares for between groups as the difference between the adjusted total sum and the adjusted within sum of squares. Thus, for the purpose of testing significance,

$$\left( B_t - \frac{A_t^2}{C_t} \right) - \left( B_w - \frac{A_w^2}{C_w} \right)$$

leads to the proper adjustment for the between sum of squares for  $X$ .

Perhaps the reader has anticipated that the  $df$ 's may change as a result of these manipulations. The new  $df$ 's are recorded in line 8 of Table 59. Note that the  $df$  for the between sum has not changed since the adjustment was not made by using the between-groups regression.

Aside from the usual methods for calculating sums of squares, we need formulas for computing sums of products in terms of raw scores. The following formulas are written for unequal  $m_j$  values, but are of course applicable for equal  $m$ 's.

$$\begin{aligned} \sum_i \sum_j (X_{ij} - \bar{X})(Y_{ij} - \bar{Y}) \\ = \sum_i \sum_j X_{ij} Y_{ij} - \frac{\sum_i \sum_j X_{ij} \sum_i \sum_j Y_{ij}}{N} \quad \text{for total} \quad (108a) \end{aligned}$$

$$\begin{aligned} \sum_i \sum_j (X_{ij} - \bar{X}_j)(Y_{ij} - \bar{Y}_j) \\ = \sum_i \sum_j X_{ij} Y_{ij} - \sum_j \frac{\sum_i \sum_j X_{ij} \sum_i \sum_j Y_{ij}}{m_j} \quad \text{for within} \quad (108b)' \end{aligned}$$

$$\begin{aligned} \sum_j m_j (\bar{X}_j - \bar{X})(\bar{Y}_j - \bar{Y}) \\ = \sum_j \frac{\sum_i \sum_j X_{ij} \sum_i \sum_j Y_{ij}}{m_j} - \frac{\sum_i \sum_j X_{ij} \sum_i \sum_j Y_{ij}}{N} \quad \text{for between} \quad (108c) \end{aligned}$$

Thus to compute the sums of products of deviations, we need the sum of all  $N$  raw score products or  $\sum_i \sum_j X_{ij} Y_{ij}$ , the sum of all the  $X$ 's or  $\sum_i \sum_j X_{ij}$ , the sum of all the  $Y$ 's or  $\sum_i \sum_j Y_{ij}$ , the sum of the  $X$ 's separately for each group or  $\sum_j \sum_i X_{ij}$ , and the sum of the  $Y$ 's

for each separate group or  $\sum Y_{ij}$ . Adding the several  $X$  sums gives the sum of all the  $X$ 's; likewise for  $Y$ 's. Note that to get the second term of (108b), or the first term of (108c), we must divide the product of the two sums for a group by its  $m$  and then sum such quotients over all  $k$  groups. The reader may find some interest in comparing formulas (108) with formulas (98), and it should be apparent that in the case of equal  $m$ 's formulas (108) can be written in the simpler way of formulas (97).

Table 60. SCORE DATA AND SUMS BASED ON RAW SCORES FOR ANALYSIS OF VARIANCE BY COVARIANCE ADJUSTMENTS

Group						
1		2		3		
Y	X	Y	X	Y	X	
14	10	11	5	7	5	$\sum \sum X = 173$
9	6	9	2	6	4	$\sum \sum Y = 268$
11	8	8	6	2	1	
12	6	10	5	10	7	$\sum \sum X^2 = 1161$
10	9	10	4	7	9	$\sum \sum Y^2 = 2642$
11	7	10	8	7	4	
11	9	12	10	6	5	$\sum \sum XY = 1688$
8	5	9	6	3	2	
11	6	10	4	2	2	$\sum (\sum X)^2 = 10,401$
12	7	11	6	9	5	$\sum (\sum Y)^2 = 25,362$
Sum	109 73	100 56	59 44	$\bar{X} = 5.77$		
Mean	10.9 7.3	10.0 5.6	5.9 4.4	$\bar{Y} = 8.93$		
$\sum Y^2$ or $\sum X^2$	1213 557	1012 358	417 246			
$\sum XY$	810	571	307			

The required computations are illustrated by using the data (fictitious) of Table 60, which contains  $Y$  and  $X$  scores for 10 cases in each of 3 groups. The scores in each of the 6 columns are separately summed to yield 109, 73, etc. The scores are squared and summed to yield 1213, 557, etc. Summing the products of the  $X$  and  $Y$  values gives 810, 571, and 307 for the 3 groups.

Summing over groups yields the double summations 173, 268, etc. Certain of these sums are then substituted into formulas (108) to secure the total, within, and between sums of products of deviations. By substituting the proper sums into formulas (97), we get the required sums of squares for the  $X$ 's and for the  $Y$ 's. Then these 3 sets of sums are entered as the first 3 rows of Table 61, which follows the pattern set forth in Table 59.

Table 61. ANALYSIS OF VARIANCE FOR  $X$  VARIABLE OF TABLE 60 BY COVARIANCE ADJUSTMENTS FOR UNCONTROLLED  $Y$

	Total	Within	Between
1. Sum of products	142.53	72.70	69.83
2. Sum of squares: $X$	163.37	120.90	42.47
3. Sum of squares: $Y$	247.87	105.80	142.07
4. $df$	29	27	2
5. Correlation	.709	.643	.912
5a. $df$ for $r$	28	26	1
6. $b_{xy}$ value	.5750	.6871	.....
7. Adjusted $\Sigma x^2$	81.42 minus	70.95 equals	10.47
8. $df$	28	26	2

Before proceeding to the covariance adjustment, let us consider the means given in Table 60. It will be noticed that the groups differ considerably on  $X$ , or the dependent variable, and that they also differ on  $Y$ , the relevant but not controlled variable. An analysis of variance based on the sum of squares for the  $X$ 's leads to a between-groups variance estimate of  $42.47/2$ , or 21.26, and a within-groups estimate of  $120.90/27$ , or 4.48. The  $F$  for testing the significance of the between-groups variance becomes  $21.26/4.48$ , or 4.75, which for the given  $df$ 's is significant at about the .02 or .03 level of significance. This analysis does not, of course, allow for the fact that the groups differ on  $Y$ . If there is correlation between  $X$  and  $Y$ , the observed differences on  $X$  may be mainly a reflection of the group differences on  $Y$ . As previously stated, the purpose of the covariance adjustment is to make statistical allowance for such uncontrolled differences.

By following the steps indicated in Table 59, we determine the values in lines 5 to 7 of Table 61. Note that the adjusted  $\Sigma x^2$  for between groups, 10.47, is secured by subtracting 70.95 from 81.42. The analysis of variance based on the adjusted sums of squares (for the  $X$ 's) gives a between-groups variance estimate of  $10.47/2$ , or 5.23, and a within-groups estimate of  $70.95/26$ , or 2.73. Then  $F = 5.23/2.73 = 1.92$ , which for 2 and 26 degrees of freedom yields a  $P$  of about .20. Accordingly, it cannot be concluded that there are significant group differences on  $X$  over and above those which would be expected because of the differences on  $Y$ .

It should be obvious that the use of the covariance adjustment method must be justified by logical and experimental considerations. When it is logical to control a variable by pairing or matching, then the covariance adjustment is defensible as a way of making proper allowance for a failure, because of infeasibility, to control the variable. The use of the covariance adjustment is not predicated on the degree of correlation between the dependent and the uncontrolled variable. If the correlation is relatively low, the adjusted values will differ but little from the unadjusted values; if high, both the total and within adjusted variances will differ considerably from the unadjusted variances, but, as we shall presently see, the extent to which the adjusted and unadjusted between-groups variances differ is not solely a function of the correlation.

It is of interest to make an actual adjustment of the  $X$  means of Table 60 for the group differences on  $Y$ . The adjustments can be made by

$$\bar{X}_{ja} = \bar{X}_j - b_{xy}(\bar{Y}_j - \bar{Y})$$

in which  $\bar{X}_{ja}$  is the adjusted value for the  $j$ th group, and  $b_{xy}$  is the *within*-groups regression coefficient. For the data of Table 60 we have

$$\bar{X}_{1a} = 7.30 - .687(10.90 - 8.93) = 5.95$$

$$\bar{X}_{2a} = 5.60 - .687(10.00 - 8.93) = 4.86$$

$$\bar{X}_{3a} = 4.40 - .687(5.90 - 8.93) = 6.48$$

Should the reader be surprised that the adjustment puts group 3 ahead, he should ponder the fact that, relative to the *within*-groups  $X$  and  $Y$  variances, the third group's  $\bar{X}$  of 4.40 was not as



far below the means of the other 2 groups as was its  $\bar{Y}$  of 5.90.

From a careful consideration of the foregoing, it will be seen that the covariance adjustment method will not necessarily reduce the differences between the means on the dependent variable. Situations arise in which groups that show marked differences on some correlated but uncontrolled variable may yield similar means on the variable being studied. Suppose that we are using 2 intact groups to investigate the relative merits of 2 learning methods, and that the initial means of the 2 groups are markedly different. We would, accordingly, expect a difference on final standing even though the 2 methods were equally efficacious. If this expected difference is not found, it follows that the method used by the group with the lower initial score was more effective in that this group overtook the other group. With groups differing on an uncontrolled variable, it is not only as proper, but also as necessary, to use the covariance technique when the groups are nearly the same on the dependent variable as when they are different. For such situations the adjustment will *increase* the between-groups variance. The adjusted variances are sometimes referred to as "reduced" variances, but it follows from the above that this term may be a misnomer for the adjusted *between*-groups variance.

The extent to which the adjusted variances lead to a level of significance different from that based on an analysis of the unadjusted values will obviously depend upon 3 things: the degree of correlation between the dependent and uncontrolled variable, the size of the differences between the groups on the uncontrolled variable, and the found differences on the dependent variable. The applicability of the covariance technique does not depend upon a minimum degree of correlation or upon a definite amount of group differences on the uncontrolled variable. But, if the within-groups correlation is low and/or there is only a small, chance difference between the groups on the uncontrolled variable, the use of the covariance adjustment may not be worth the effort. Obviously, if a variable correlates near zero with the dependent variable, it need not be controlled experimentally or statistically.

The covariance method can be extended to make adjustments for group differences on more than 1 uncontrolled variable. This involves the use of multiple regression, but computationally it is perhaps simpler to handle the adjustments in terms of multiple

$r$ 's. We need 2 multiple correlation coefficients, one obtained by way of correlations based on within-groups sums of squares and of products, and the other by way of correlations based on total sums of squares and of products.

If, for example, allowance is to be made for 3 uncontrolled variables,  $Y_1$ ,  $Y_2$ , and  $Y_3$ , we will need 6 (one for each pair of variables— $X$  is the fourth or dependent variable) auxiliary tables consisting of entries like those in lines 1, 2, and 3 under the "total" and the "within" columns of Table 59 (or Table 61). We can then calculate 2 sets of intercorrelations (each auxiliary table will lead to 2  $r$ 's when the substitutions called for in line 5 of Table 59 are made) among the 4 variables, and from these we compute, by the methods set forth in Chapter 11, two  $r^2_{x \cdot y_1 y_2 y_3}$  values. Let us designate the multiple based on the total sums as  $R_t$  and that based on the within sums as  $R_w$ .

With these 2 multiple  $r$ 's available, we may rewrite line 7 of Table 59 as

$$B_t(1 - R_t^2) \text{ minus } B_w(1 - R_w^2) \text{ equals adjusted } B_b$$

with respective  $df$ 's of

$$N - n, \quad N - k - (n - 1), \quad k - 1$$

for the  $n$  variable problem (1 dependent, plus the number of uncontrolled variables included in the adjustments).

Remark: The use of the covariance adjustment technique is far superior to attempts at pairing individuals from the intact groups on the basis of 1 or more uncontrolled variables, a procedure which inevitably leads to a reduction of sample size and also runs astride a regression difficulty.\*

**Evaluation of changes.** In Chapter 6 (pp. 90-92) we discussed the usually advocated method for comparing changes shown by experimental and control groups (applicable also for 2 experimental groups). We have, with  $i$  and  $f$  standing for the pretest and posttest measures and  $E$  and  $C$  standing for experimental and control groups,

$$D = \bar{D}_E - \bar{D}_C = (\bar{X}_{fE} - \bar{X}_{iE}) - (\bar{X}_{fC} - \bar{X}_{iC})$$

as the net change, the change shown by the experimentals cor-

\* See Thorndike, R. L., Regression fallacies in the matched groups experiment, *Psychometrika*, 1942, 7, 85-102.

rected for that shown by the controls. We may rearrange the  $\bar{X}$ 's, yet maintain the numerical value of  $D = \bar{D}_E - \bar{D}_C$ , as follows:

$$D = (\bar{X}_{fE} - \bar{X}_{fC}) - (\bar{X}_{iE} - \bar{X}_{iC})$$

from which it is seen that the net change may also be thought of as the final difference between the 2 groups corrected for their initial difference. Such a correction involves the assumption that each unit of difference in initial standing will produce a unit of difference in final standing. In other words, this type of adjustment implies a 1-to-1 relationship between initial and final scores. Since a perfect correlation is never found or approached in practice, one may question whether the usual procedure of comparing changes is really defensible.

It is, of course, entirely logical that group differences on final scores, which we may here call the dependent variable, should be corrected for group differences on initial standing as an uncontrolled variable. The covariance adjustment technique provides a way of correcting final means for initial differences, with due allowance for the *degree* of correlation between initial and final scores. The ordinary and the covariance method differ not only in the correction but also in the resultant sampling error. The ordinary technique uses a standard error which definitely includes, either explicitly or implicitly, the variance for both initial and final scores and the correlation of initial with final, whereas the error term used in the covariance method is a direct function of the degree of correlation and of the variance for the final scores only. In other words, the net differences being tested are not the same, and neither are the error terms the same. The covariance method will, in general, be more sensitive. The student should read Professor R. A. Fisher's discussion on this point.†

† Chapter IX in Fisher, R. A., *Design of experiments*, London: Oliver and Boyd.

## Distribution-Free Methods

The tests of significance involving  $F$ ,  $t$ , or  $CR$  (critical ratio) are based on an assumption of normality. For large samples the degree of skewness that can be tolerated is a function of sample size (see p. 100 for effect of skewness on sampling distribution of the mean), but for small samples skewness becomes a disturber. Occasionally psychologists have score data which are so markedly skewed in distribution (e.g., certain scoring categories of the Rorschach test) that it is not possible to normalize the distribution either by McCall's  $T$  scaling technique (p. 39) or by mathematical transformations.\* Accordingly, we may need what have been called nonparametric or, more appropriately named, distribution-free methods.

Actually, the  $\chi^2$  technique can be classified as distribution-free—no assumptions are made about the distribution of the variable or variables underlying the categories. Likewise, tests of the significance of relationships by way of Spearman's rho or Kendall's tau (pp. 208-210) do not depend on assumptions regarding trait distribution.

In general, distribution-free methods, when applied for comparative purposes to data which are normal, are not as sensitive (that is, as powerful for avoiding type II errors) as the appropriate  $CR$ ,  $t$ , or  $F$  technique. Consequently, it is unwise to use a nonparametric method as a short-cut for testing significance when the assumption of normality is tenable.

**The sign test.** Perhaps the simplest of all distribution-free methods is the "sign" test which is applicable for testing the dif-

\* See Mueller, C. G., Numerical transformations in the analysis of experimental data, *Psychol. Bull.*, 1949, **46**, 198-223.

ference between 2 correlated sets of scores. The procedure is to consider the  $N$  pairs of differences,  $X_1 - X_2$ , some of which will be plus, some minus (with an occasional zero). If there is no difference between the 2 sets of scores we would expect the plus and minus signs to be equally divided. To test whether there are more plus signs than reasonable on a chance basis, the binomial,  $(p + q)^N$  with  $p = .50$ , is used ( $N$  is for the pair differences having a sign; it is the sample size less the number of zero differences) in the manner discussed earlier (pp. 49-51). For effective  $N$  larger than 10 we may use either the normal curve approximation to the binomial (pp. 46-49) or the  $\chi^2$  approximation (pp. 212-214). Whether one uses the binomial itself or one of the approximations, care must be taken to secure a  $P$  that represents whichever—a one-tailed or a two-tailed—test is appropriate for the hypothesis being tested.

**The "median" test.** A procedure for testing the difference between 2 sets of *independent* scores is to use the median for the 2 groups combined as a basis for dichotomizing. This leads to a fourfold table: above vs. below (the median) on 1 axis, group vs. group on the other. Then the  $\chi^2$  test for the fourfold table (pp. 224-225) may be employed, with Yates's correction if necessary. With very small  $N$ 's the exact probability method (pp. 240-242) would be used. The idea back of the median test is simply that 2 samples drawn from 2 populations having the same median should yield equal splits. In practice, difficulties are sometimes encountered in attempting to dichotomize exactly at the median. When the median is an integer and several scores are equal to the median, the dichotomy can be taken as those scores which *exceed* the median vs. those which do not exceed the median.

**Median test for more than 2 independent groups.** This is a straightforward extension of the median test to provide an over-all test of the differences between, say,  $C$  independently drawn groups. On the basis of the median of the distribution of the  $C$  groups combined, the scores are dichotomized (as near the median as possible). This will lead to a 2 by  $C$  table from which one may obtain a  $\chi^2$  with  $C - 1$  degrees of freedom.

Whether we are dealing with 2 groups or with  $C$  groups, the  $N$ 's for the groups need not be equal for use of the median test.

**Test of  $C$  correlated sets.** Suppose  $R$  individuals (or  $R$  sets of matched persons) with scores under  $C$  different conditions. This is, of course, the familiar 2-way classification setup which we dis-



cussed (Chapter 16) under the analysis of variance, mixed model. We shall be concerned here with testing the effect of the  $C$  conditions.

For the distribution-free test we need to arrange the scores in  $R$  rows and  $C$  columns. The median score for each row is determined, and then the scores in each row which *exceed* the row median are assigned a plus. This will lead to  $C/2$  pluses in each row if  $C$  is an even number, and to  $(C - 1)/2$  if  $C$  is an odd number. Any row having all scores identical is ignored, and therefore  $R$  in the following exposition is the original  $R$  minus the number of rows with identical scores in the row. The pluses for each column are counted. Let  $T_c$  stand for the number of pluses in the  $c$ th column. For  $C$  even we will have a total of  $RC/2$  plus signs. If there are no real column effects we would expect these  $RC/2$  pluses to be distributed evenly over the columns. Thus we would expect  $R/2$  pluses per column, on the basis of the null hypothesis. For  $C$  odd we will have a total of  $R(C - 1)/2$  plus signs, which when divided evenly among the columns would give  $\frac{R(C - 1)}{2C}$  as the chance expected number of pluses per column.

With an observed number of pluses per column, and an expected number per column, we have what begins to look like a  $\chi^2$  situation, and so it is, but not an ordinary one. The manner of assigning pluses leads to subtle restrictions, such that we have

$$\chi^2 = \frac{C(C - 1)}{RA(C - A)} \sum (T_c - RA/C)^2 \quad (109)$$

with  $C - 1$  degrees of freedom. The value of  $A$  is taken as  $C/2$  when  $C$  is even and as  $(C - 1)/2$  when  $C$  is odd. Note that  $RA/C$  is the expected value on the assumption of no real column effects.

Mood,<sup>†</sup> who presents the foregoing test, states that for the test to be valid either  $R$  should be as large as 10 or  $RC$  should be 20 or more. When  $C = 2$ , this method yields exactly the same  $\chi^2$  as that obtained by the  $\chi^2$  technique applied to the sign test.

**Mann-Whitney *U* test.** This test, which is applicable only to results based on 2 independent groups, involves rank ordering the scores, for the 2 groups combined, from *greatest* (rank 1) to *least* (for which the rank will be  $N = N_1 + N_2$  unless there are ties

<sup>†</sup> See Chapter 16 of Mood, A. M., *Introduction to the theory of statistics*, New York: McGraw-Hill, 1950.



for the bottom position). When ties occur, each person involved is assigned the average of the ranks that would be assigned in case the tied persons could be differentiated (see p. 209). Then the ranks so assigned are summed separately for each group. Let  $T_1$  and  $T_2$  represent these 2 sums. (As a check on the arithmetic,  $T_1 + T_2$  should equal  $\frac{N(N+1)}{2}$ , the sum of the first  $N$  natural numbers.)

When both  $N_1$  and  $N_2$  are 8 or greater, the statistic

$$U_1 = N_1N_2 + \frac{N_1(N_1 + 1)}{2} - T_1 \quad (110)$$

is distributed normally about a chance expected value, or mean, given by  $N_1N_2/2$ , and with variance of  $N_1N_2(N_1 + N_2 + 1)/12$ . We then have

$$\frac{x}{\sigma} = \frac{U_1 - N_1N_2/2}{\sqrt{\frac{N_1N_2(N_1 + N_2 + 1)}{12}}}$$

as a unit normal deviate by which the significance of  $U$  as a deviation from the null hypothesis expected value is determined. If, as an alternate, we define  $U$  by replacing  $T_1$  with  $T_2$  and  $N_1$  with  $N_2$  (in the second term), we will have  $U_2$ . Now  $U_1$  and  $U_2$  will deviate to the same extent, but in opposite directions, from  $N_1N_2/2$ .

When  $U_1$  is larger than  $N_1N_2/2$ , the direction of the difference between the 2 sets of scores is such that group 1 is superior to group 2. (If ranks are assigned with the least score as rank 1, and so on, the value of  $U_1$  will be smaller than  $N_1N_2/2$  when group 1 is superior.) For  $N_1$  and  $N_2$  less than 8, special tables are required for judging the significance of  $U$ . These may be found in an article by Mann and Whitney, *Annals math. Stat.*, 1947, **18**, 50-60. Note: these tables are set up on the basis of the least score being assigned a rank of 1.

The  $U$  test is more powerful than the median test, and hence is preferable to the latter unless there are too many ties in ranks.

For other distribution-free methods the reader is referred to an article by Lincoln Moses, *Psychol. Bull.*, 1952, **49**, 122-143. It might be remarked that at present little is known about the relative merits of the many available techniques.

## CHAPTER 19

### Remarks on Error Reduction

In this brief chapter we shall attempt a summary and integration of implications, scattered through this book, having to do with the reduction of error variance in psychological research. In a sense, this is an extension of an earlier discussion (pp. 88-90). Some of the additional concepts and techniques could not have been introduced at that time since an understanding of them is dependent on material presented in the intervening chapters.

For our present purpose, we shall subsume errors under 3 headings: measurement or observational errors, errors in inferring population parameters in field or survey studies, and errors in experimental testing of hypotheses. About the first of these, we remark only that errors of measurement can be reduced by developing more reliable tests or (when feasible) by averaging repeated measurements.

#### FIELD STUDIES (SURVEYS)

Surveys for the purpose of gauging opinion, and studies designed to establish normative data, require large scale sampling. The aim is to secure a sample which is unbiased, that is, representative of a defined population, with chance sampling errors as small as possible. We shall limit ourselves to 3 sampling methods: random, stratified, and area.

**Random sampling.** The conditions of random sampling have been specified earlier (p. 55). By the method of random sampling it is fairly easy to arrive at a representative sample, provided the universe has been catalogued. Thus, if one wishes a

representative sample of school children of a certain grade in a city, he can secure it by a purely mechanical scheme, such as taking every  $n$ th card from the files. Although this type of *systematic* sampling does not exactly satisfy the conditions of random sampling, it will assure a random sample unless the cards have been systematically arranged (in a somewhat peculiar order). The use of the random method for sampling an uncatalogued population involves so many difficulties in psychological research that no schemes are to be found in the literature.

Increasing sample size is the only way by which one can reduce chance errors when the random method is being employed. That sheer sample size is not enough to reduce nonrandom errors is evidenced by the *Literary Digest* straw polls, which rested on the assumption that the population of telephone subscribers and car owners was not different in its voting preference from the entire population of potential voters. This happened to hold before 1936, so that replies to ballots mailed at random to telephone subscribers and car owners forecasted fairly accurately the election results. Despite a very large sample, the *Digest* poll failed miserably in 1936; this failure is attributed to the alignment of voting to income levels, an alignment that did not exist in prior years.

**Stratified sampling.** In the stratified method, one or more individuals are pulled at random from each of several strata, the number in the sample from each stratum being proportional to the universe number in the stratum, and the strata are predetermined by knowledge of some control variable or variables. Psychologists who sample so as to secure proportionate representation from the several occupational levels are, in effect, using the principle of stratification. It should be obvious that the method can be used only when information is available on some variable or variables which permits their use in setting up the strata, and when cases within the strata can be drawn randomly.

When the sampling is for attributes by the stratified method, the standard error of an obtained proportion,  $P$ , is given, in terms of information yielded by the sample, approximately by

$$\sigma_P = \sqrt{\frac{PQ}{N} - \frac{\sigma_p^2}{N}} \quad (111)$$

where  $P$  equals the proportion in the total sample,  $N$ , who possess

the attribute,  $Q = 1 - P$ , and  $\sigma_p^2$  is the weighted variance of the several strata proportions about the sample value  $P$ . A casual examination of formula (111) shows that the magnitude of the error is less for a stratified sample than for a random sample, and that the increase in precision depends upon one's ability to stratify the universe in such a way as to secure strata which are really different with regard to the attribute being studied.

For stratified sampling, the variance of the mean may be written as

$$\sigma_{\bar{x}}^2 = \frac{1}{N} (\sigma^2 - \sigma_{\bar{x}_r}^2) \quad (112)$$

where  $\bar{X}$  = the sample mean,  $\sigma^2$  = the sample variance, and  $\sigma_{\bar{x}_r}^2$  = the weighted variance of the means of the several strata about the total sample mean. If stratification has been accomplished by use of a variable,  $Y$ , which is linearly related to the variable being studied, the formula can be written in the form

$$\sigma_{\bar{x}}^2 = \frac{1}{N} (\sigma_x^2 - \sigma_x^2 r_{xy}^2) \quad (113)$$

It will be noticed that stratified sampling does lead to greater precision in the sense of smaller chance error, but only when the control or stratifying variable is related to the variable being studied.

The *quota* method involves the use of strata, but selection within the strata is not done on a random basis—the field worker merely fills a quota by securing the correct proportion per strata; selective factors leading to bias can easily operate.

**Area sampling.** There is considerable evidence that area or “pin point” sampling is the best method yet devised for drawing samples in survey studies. Its use, however, depends on the availability of extensive facilities. The student who is interested in this, or the stratified, method will wish to turn to detailed treatments of the subject.\*

## SAMPLING ERRORS IN EXPERIMENTATION

The formation of groups for experimental purposes can be accomplished (1) by random sampling—the random assigning of

\* Yates, F., *Sampling methods for censuses and surveys*, New York: Hafner, 1949; Deming, W. E., *Some theory of sampling*, New York: Wiley, 1950.

individuals to the groups, (2) by pairing, (3) by using sibs or littermates, (4) by matching distributions, and (5) by using the same person under all the experimental conditions. The last mentioned will not be feasible when practice or fatigue effects are likely.

For methods 2, 3, and 5 the statistical analysis is by way of the analysis of variance (mixed model) with rows standing for the matched persons or litters or individuals, respectively, for the 3 methods. The  $F$  test of the significance of the differences among the correlated means (the means for conditions) involves an error term which is freed of the row variation; stated differently, the error term (an estimate of a 2-way interaction variance) tends to be small if the correlations between the matched persons or between sibs or between scores on the same persons are large. The foregoing argument holds, of course, for just 2 experimental groups (or an experimental and a control group) as well as for 3 or more groups.

Thus, compared to method 1 (random assignment), greater precision is attainable by using method 2 or 3 or 5. Before discussing method 4, let us again consider the situation where groups are needed for just 2 conditions. If the groups are formed by pairing individuals, the sampling variance of the difference between the 2 means is, as we learned in Chapter 6, given by

$$\sigma_D^2 = \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2} - 2r_{12}\sigma_{\bar{X}_1}\sigma_{\bar{X}_2} \quad (25b)$$

The gain in pairing, over random assignment, depends on the magnitude of  $r_{12}$ . It can be shown that if the pairing is done on the basis of variable  $Y$ , the value of  $r_{12}$  will be  $r^2_{xy}$ , and in case 2 or more variables are controlled by pairing,  $r_{12}$  will be the square of the multiple correlation between the dependent variable,  $X$ , and the control variables. The reason for pairing, it will be recalled, is to make the groups comparable on certain variables which might affect the outcome of the experiment. We now see explicitly that the advantage of pairing depends definitely on how highly the variables, so controlled, are correlated with the dependent variable. No correlation, no gain; low correlation, little gain.

Method 4 is another way of making groups comparable on pertinent variables. Instead of pairing persons, distributions are matched for the  $Y$  variable, to be controlled, in such a manner



that the 2 groups contain the same proportions of cases in the several intervals as hold for a supply distribution on  $Y$ . The sampling variance of the difference between the 2  $X$  means is given by

$$\sigma_D^2 = \sigma_{X_1}^2(1 - r_{xy}^2) + \sigma_{X_2}^2(1 - r_{xy}^2) \quad (114)$$

If the matching has been made on the basis of several control variables, the 2 correlations (1 for each group) become the multiple  $r$ 's between  $X$  and the control variables.

From (114) one may deduce the following fact: Where 2 groups have been separately matched as to distribution on the same control variable(s), the standard error of the difference can be obtained without the restriction of the ordinary pairing procedure, which requires that there be an equal number of cases in the 2 groups. The reader will note that either term in formula (114) is, as might be expected, identical to formula (113) for the sampling variance of a mean when the stratified method is used. The method of matching distributions is particularly useful when the cost per case is much greater in the experimental group than in the control group. Precision can be increased by taking a larger control group—a possibility also when the groups are chosen by randomization.

The use of paired individuals for experimental (and control) conditions has long been recognized as a sound procedure. One might argue, however, that the advantages of pairing have been overstressed. The gain in error reduction may not be appreciable. The advocates of pairing say that they are not willing to risk randomization as a method for setting up groups, but it should be noted that there are always numerous variables which might affect the outcome of an experiment that are never controlled except by randomization. Thus one can seldom, if ever, completely avoid placing faith in the randomization process. Random differences between groups never have more than a random effect on the results; the error formulas always include all random effects. When pairing leads to only a slight reduction in error, we have evidence that the pairing procedure may not have been worth the effort involved.

It should be noted that an original group which is split into experimental groups either by the random method or by pairing must be regarded as representative of some defined universe, and



that such conclusions as are drawn from the experiment cannot be generalized unless it can be shown that the defined universe is representative of the generality of mankind with respect to the variables being studied. In other words, those who use the college sophomore as a laboratory representative of mankind have not avoided, by showing that selective factors did not render their experimental groups noncomparable, the necessity of bridging the gap between the sophomore's behavior and that of the typical human being.

At this point, we remind the student that the covariance adjustment method (Chapter 17) is an entirely legitimate technique for allowing for uncontrolled variables and at the same time reducing error variance.

It is appropriate to end this discussion (and the text) with an example of an experiment in which error reduction might have been achieved by judicious planning. The Lanarkshire milk experiment in England involved the daily feeding of three-fourths of a pint of raw milk to 5000 children and of an equal amount of pasteurized milk to another group of 5000 over a period of 4 months. These 10,000, plus a control group of 10,000, were measured for height and weight at the beginning and the end of the 4-month period. Since the purpose of the experiment was to check on the relative merits of raw vs. pasteurized milk, the control group was nonessential. (It is an interesting commentary on the magic of the word "control" that very frequently a control group is used when not needed.) Despite large numbers, the feeder and control groups were not comparable as regards initial height and weight, the operating selective factor being the benevolent attitude of school teachers who apparently thought the research would not be harmed if preference was given frail, undernourished children in choosing individuals for the feeder groups. Either a carefully supervised random, or a definite pairing, procedure would have avoided this selective bias, but what is more important and more relevant to our present topic is the claim in a paper † by "Student," so far not refuted, that the use of 50 pairs of identical twins would have yielded as precise information at only 2 per cent of the cost of the original experiment, or at a saving of approximately 35,000 prewar dollars.

† "Student," The Lanarkshire milk experiment, *Biometrika*, 1931, 23, 398-406.

# EXERCISE MATERIAL FOR ELEMENTARY STATISTICS\*

## Chapter 2

1. *a.* Make separate frequency distributions for the marks of the two groups of students in Table I. Use intervals of size 5.
- b.* Determine also the cumulative frequencies for each group.

Table I. FINAL EXAMINATION MARKS FOR A CLASS IN STATISTICS

Students with No Calculus ( $N = 36$ )						Students with Some Calculus ( $N = 22$ )			
103	150	139	79	150	134	137	139	112	139
98	79	94	137	118	113	151	124	80	153
106	93	106	137	91	109	131	94	96	77
71	101	92	74	106	87	133	123	101	115
108	113	103	108	114	105	115	90	154	122
120	95	83	93	109	97	111	135		

2. *a.* Make separate frequency distributions for the two groups of scores in Table II. Use intervals of size 3.
- b.* Determine also the cumulative frequencies for each group.

Table II. SCORES ON FINAL EXAMINATION FOR A COURSE ON PSYCHOLOGICAL TESTS

Undergraduates ( $N = 32$ )						Graduate Students ( $N = 23$ )			
70	72	76	66	76	80	84	80	90	82
67	69	90	50	76	47	79	62	77	89
51	58	71	88	65	54	73	74	87	76
89	64	80	67	71	90	85	95	78	69
91	71	63	81	87	81	78	86	92	
79	79								

3. *a.* Draw a frequency polygon for the distribution in Table III, part A.
- b.* Draw an ogive for the data of Table III, part A.
4. *a.* Draw a frequency polygon for the distribution of Table III, part B.
- b.* Draw an ogive for the data of Table III, part B.

\* These exercises are so arranged that, in general, each even-numbered exercise is of the same type as its immediately preceding odd-numbered exercise.

## Chapter 3

5. For the scores of Table I, compute separately for the two groups:
  - a. the medians, using the undistributed scores.
  - b. the medians, using the frequency distributions.
6. Repeat exercise (5) with the data of Table II.
7. Compute the mean for each group in Table I by
  - a. the definition formula for the mean.
  - b. the arbitrary origin method.
8. Repeat exercise (7) with the data of Table II.
9. Combine the two distributions for the data of Table I, compute the mean by the arbitrary origin method, and check by using the formula for securing the mean for a combined group (use the means obtained by the arbitrary origin method for this check).
10. Repeat exercise (9) with the data of Table II.

Table III. DISTRIBUTIONS OF IQ'S, FORM L OF REVISED STANFORD-BINET SCALE

IQ	A. Ages 2½-5½		B. Ages 6-13	
	<i>f</i>	<i>cu f</i>	<i>f</i>	<i>cu f</i>
170-179			1	1623
160-169	4	728	1	1622
150-159	4	724	3	1621
140-149	11	720	29	1618
130-139	41	709	73	1589
120-129	82	668	140	1516
110-119	175	586	308	1376
100-109	193	411	407	1068
90-99	107	218	335	661
80-89	76	111	215	326
70-79	20	35	76	111
60-69	7	15	30	35
50-59	5	8	4	5
40-49	2	3	1	1
30-39	1	1		
<i>N</i> = 728			<i>N</i> = 1623	

11. Suppose the following intervals for the heights of trees, measured to the nearest inch: 250-274; 275-299; 300-324.  
 What is the midpoint of the 250-274 interval?  
 What is the lower limit of the 275-299 interval?  
 What is the upper limit of the 275-299 interval?
12. Given the following intervals for length, to the nearest inch: 42-44; 45-47; 48-50.  
 What is the midpoint of the 42-44 interval?

What is the lower limit of the 48-50 interval?

What is the upper limit of the 45-47 interval?

13. Specify the midpoint values for the bottom interval of each of the following:

Age at Last Birthday	Weight to Nearest Pound	Size of College Classes
30-34	144-147	15-19
25-29	140-143	10-14

14. Give the midpoints of the first of the following intervals:
- 45-49; 50-54 (age at last birthday).
  - 45-49; 50-54 (size of college classes).
  - 45-49; 50-54 (length to nearest centimeter).
15. What arbitrary rule would you follow in determining lower limits and midpoints of intervals for the data of Tables I and II?
16. As is commonly known, an IQ is obtained by multiplying the quotient, MA/CA, by 100 and rounding to nearest integer. Now CA is taken as age to the *nearest* month, whereas an MA of, say, 88 months means *at least* 88 months. Considering these facts, what would be the *exact* lower limit for the bottom interval of Table III?
17. Compute the median,  $Q_1$ , and  $Q_3$  for the distribution in Table III, part A.
18. Compute the median,  $Q_1$ , and  $Q_3$  for the distribution in Table III, part B.
19. Compute the 10th and 90th percentile points for the distribution in Table III, part A.
20. Compute the 20th and the 80th percentile points for the distribution in Table III, part B.
21. Using the results of exercises (17) and (19), locate the five points,  $Q_1$ , the median,  $Q_3$ ,  $P_{10}$ , and  $P_{90}$ , on the base line of your ogive curve for the distribution of Table III, part A. Divide the ordinate on the right-hand side (the ordinate at  $IQ = 170$ ) into approximate fourths. Draw a line from each of the five base-line points up to the ogive, then horizontally to the right. Notice where these horizontal lines hit the ordinate on the right-hand side.
22. Using the results of exercises (18) and (20), locate the five points,  $Q_1$ , the median,  $Q_3$ ,  $P_{20}$ , and  $P_{80}$ , on the base line of your ogive curve for the distribution of Table III, part B. Divide the ordinate on the right-hand side (the ordinate at  $IQ = 180$ ) into approximate fourths. Draw a line from each of the five base-line points up to the ogive, then horizontally to the right. Note where these horizontals hit the ordinate on the right-hand side.
23. a. Compute the SD's ( $\sigma$ 's) for the two groups in Table I (use arbitrary origin method).  
b. Combine the two distributions in Table I, compute the SD, then see whether the obtained SD agrees with that secured by using formula (8).
24. Repeat exercise (23) with the data of Table II.
25. For the distribution of IQ's in Table III, part A, the mean is 106.68 and the standard deviation is 17.41.

- a. Determine the two points defined by  $M \pm \sigma$ .
  - b. Determine the two points defined by  $M \pm 2\sigma$ .
  - c. Locate these four points, also the mean, on the base line of your frequency polygon for the data of Table III, part A. Erect ordinates from each of these five base-line points to the polygon, and study the resulting picture.
  - d. Determine approximately the percentage of cases between  $M \pm \sigma$ ; also between  $M \pm 2\sigma$ .
26. The distribution of IQ's in Table III, part B, has a mean of 103.34 and a  $\sigma$  of 16.88. Repeat exercise (25), using the values and polygon for the data of Table III, part B.
27. Suppose the mean score on a statistics quiz is 35, the median is 36, the  $SD$  is 6, and the quartile deviation is 4.
- a. If to each person's score we added 50 points, what values would we then get for the mean, the median, the  $SD$ , and the quartile deviation?
  - b. If we doubled each person's score, what would be the values of the new mean and  $SD$ ?
28. Given that the distribution of scores on a quiz leads to a mean of 40, a median of 38, an  $SD$  of 9, and a quartile deviation of 6.
- a. If we added 10 points to the scores of each student, what would be the values for  $M$ ,  $Mdn$ ,  $SD$ , and  $Q$ ?
  - b. If all scores were halved, what would be the values of the mean and the  $SD$ ?
29. For each of the following three sets of two groups, determine the mean for the two groups combined.
- a.  $N_1 = 40$ ,  $M_1 = 28$ ;  
 $N_2 = 40$ ,  $M_2 = 23$ .
  - b.  $N_1 = 100$ ,  $M_1 = 44$ ;  
 $N_2 = 60$ ,  $M_2 = 60$ .
  - c.  $N_1 = 12,489$ ,  $M_1 = 228.63$ ;  
 $N_2 = 6971$ ,  $M_2 = 228.63$ .
30. Given that the mean weekly pay of the seven working members of the Jones family is \$55 and the median is \$50 (both after deductions).
- a. What is the weekly "take home" of the family?
  - b. Suppose that Daddy Jones, already the best paid, receives an increase which after deductions amounts to \$6 a week. What is the new mean? What is the new median?
31. If an  $SD$  is 9 when computed from a frequency distribution with intervals of size 6, what would you expect it to be if computed by using the definition formula for  $SD$ ?
32. How large is the grouping error in an  $SD$  of 13 computed from a distribution with intervals of size 12?

## Chapter 4

33. Assume that the IQ's for a large number of unselected elementary school children are distributed as a normal curve with a mean of 100 and an *SD* of 17.
- The first quartile point will be near what value?
  - The percentage with IQ's above 130 will be?
  - The middle 80 per cent will fall between what values?
  - The 99th percentile will be near what IQ value?
  - The percentage with IQ's below 70 will be?
34. Let us presume that the Army General Classification Test yields a normal distribution of scores, with mean of 100 and *SD* of 20.
- The value of the third quartile will be near what score?
  - The first percentile point will be at what score?
  - Between a score of 70 and a score of 130 will be found what percentage of the cases?
  - The middle 60 per cent of scores will fall between what score values?
  - The value of the quartile deviation will be what?
35. One way to comprehend the meaning of either sizable or small differences between groups is to consider the extent to which the distributions overlap. Given the following data for weights of college students:

Men:  $M = 142$ ,  $\sigma = 15$ ; Women:  $M = 120$ ,  $\sigma = 12$

- Assuming normality for both distributions, how many men per thousand are lighter than the average woman? Determine the number of women per thousand who are heavier than the average man.
36. If the mean height for college men is 68.5 inches and the *SD* is 2.8, and if the mean height for college women is 64.5 and the *SD* is 2.5, what proportion of women exceed the average man in height? What proportion of men fall below the average height for women?
37. If the IQ's of children of professional people average 116, with a  $\sigma$  of 12, what percentage of such children would you expect to fall below 100, the general average (assume normality)?
38. Using the data of exercise (36), determine the percentage of men who are more than 6 feet tall, and the percentage of women who are shorter than 5 feet.
39. Suppose that the distribution of numerical grades in a course is normal with a mean of 60 and an *SD* of 10. The instructor wishes to assign letter grades as follows: 15 per cent A's, 35 per cent B's, 35 per cent C's, and 15 per cent D's. Determine to the nearest score the dividing line between the A's and B's, between the B's and C's, and between the C's and D's.
40. Suppose that it has been decided to use a five-letter grading system, A, B, C, D, and E, and that it is required that the letters shall correspond to "equal" distances on the base line, the whole of which is taken to be six sigmas. Assuming normality, what percentage would be assigned A's; B's; C's?



41. Determine the height of the unit normal curve at the point which is 1.2 sigma units below the median; at the third quartile point; at the point which is two  $Q$ 's below the median.
42. What is the height of the ordinate of the unit normal curve corresponding to the  $x/\sigma$  value that cuts off the upper 10 per cent of the curve? The lower 25 per cent? The upper 2.5 per cent?
43. Frequently, one must be able to translate percentile scores to standard scores and vice versa (assume normality).
  - a. What are the standard scores (to the nearest tenth) which correspond to the following percentiles: 67th, 44th, 20th, 99th?
  - b. What are the percentile equivalents (to nearest value) of the following standard scores: +1.04, -1.34, -1.75, +2.06?
44. Suppose a typical bell-shaped distribution:
  - a. What is the approximate percentile value of the following points: the mean,  $Q_3$ , the point which is one  $SD$  above the mean, the first decile point?
  - b. What is the percentile value of an IQ of 140? 80? [See exercise (26) for needed mean and  $SD$ .]
45. What is the sigma distance between the following (assume normality):
  - a. the 10th and the 90th percentile points?
  - b. the 25th and the 75th percentile points?
  - c. the 1st and the 99th percentile points?
46. If a distribution of scores is normal, what is the sigma distance between the 10th and 20th percentile points? between the 40th and 50th percentile points?
47. Given that a reading test for unselected 10-year-olds yields a mean of 50 and an  $SD$  of 10, while an arithmetic test gives a mean of 48 and an  $SD$  of 8. If Joe Bloke scores 52 on reading and 50 on arithmetic, is he better in reading than in arithmetic? Why?
48. If a student's reading rate score falls at the 20th percentile, and his standard score on reading comprehension is -1.4, would you conclude that his comprehension was superior to his rate? Why?
49. If the scores of a positively skewed distribution are converted to percentile scores, what will be the form of the distribution of percentile scores?
50. If the scores of a skewed distribution were converted to standard scores, what would be the form of the resulting distribution of standard scores?

## Chapter 5

51. If you tossed four unbiased pennies 160 times, how often would you expect to have two heads and two tails?
52. Suppose you roll a pair of fair dice once. What is the probability that exactly eleven spots will turn up?
53. Suppose that you are rolling two fair dice, one red and the other white. What is the probability of obtaining a three spot on the red die and a four spot on the white one?

54. In that back-alley game known as "crap shooting," the obtaining of spots on the 2 dice totalling 7 seems to be of paramount importance at certain times. What is the probability of rolling a 7 (assume gentlemen's dice)?
55. Suppose that we have 3 pyramidal objects (perfectly homogeneous) which can be rolled like dice. The sides of each are numbered 1, 2, 3, 4; and success is defined as the getting of 4's on the down sides. Determine the probability for obtaining exactly three 4's; exactly two 4's; exactly one 4; and no 4's. What is the probability of securing at least two 4's?
56. If you were dealt 1 card from each of 5 well-shuffled decks, what is the probability of all 5 cards being spades?
57. The probability of drawing a red card from an ordinary (and well-shuffled) deck is  $\frac{1}{2}$  and the probability of drawing a heart is  $\frac{1}{4}$ . Why isn't  $\frac{1}{2}$  plus  $\frac{1}{4}$  the probability of drawing either a heart or a red card, or is it?
58. Suppose that for a class of 100 the number of A's given on the first quiz is 15 and that the number of A's on the second quiz is also 15. Suppose further that the names of the students are placed on slips which are then well mixed in a hat. We might say that the probability is .15 that a name drawn from the hat will be that of a student who received an A on the first quiz; likewise, the second quiz. Why might it be erroneous to say that the probability is .15 times .15 that the drawn name belongs to a student who made A's on both quizzes?
59. Suppose a true-false quiz of 6 questions; what is the probability of securing a perfect score on a chance (or pure guessing) basis?
60. Some folks have pointed out that the blindfold test of the ability to distinguish brands of cigarettes is befuddled with guessing. Suppose that you are to test a friend who claims that he can tell the difference between Luckies and Camels. At 5-minute intervals you present him with a cigarette, either a Camel or a Lucky according to the flip of a coin, until he has tried 8 cigarettes. What is the probability that he would by chance alone name all 8 cigarettes correctly?
61.
  - a. Toss 6 coins 64 times; for each toss tally the number of heads that turn up, thereby obtaining a frequency distribution with an  $N$  of 64. Label this Series  $A$ . Toss the coins 64 more times, and label the resulting distribution as Series  $B$ . Then combine the 2 distributions.
  - b. Using the binomial expansion, ascertain the expected distribution when 6 coins are tossed 64 times; 128 times.
  - c. Compute the mean and standard deviation for each of your 3 distributions; also for the expected distribution (round to 2 decimals).
  - d. Determine the proportion of times that 3 heads, also 6 heads, turned up in each series, and in the combined series. Compare these results with the expected proportions.
  - e. Subtract the mean of Series  $A$  from that of Series  $B$  (keep sign if negative). For the proportion of times 3 heads turned up, subtract the Series  $A$  proportion from that for Series  $B$  (keep sign).
  - f. Bring all the results to class so that frequency distributions may be made for  $M$ 's,  $\sigma$ 's, proportions, and differences between  $M$ 's and between proportions.

62. Do exercise (61), using 7 coins.
63. If 42 of 60 rats turn to the right at the first choice point in a maze, would you conclude that rats, in general, prefer to turn to the right at this choice point? First get your answer by using the appropriate standard error, then check by using the binomial expansion and normal curve approximation thereto.
64. If at a particular time 50 per cent of all eligible voters favor the Democrats, how often would polls based upon random samples of size 400 yield percentages of 55 or over as favoring Democrats?
65. Items on an intelligence test of the Binet type are at times assigned an index of difficulty which is nothing more than the percentage passing the item. Given the following for an item: of 100 12-year-olds, 60 per cent passed; of 100 13-year-olds, 80 per cent passed. When possible sampling errors are considered, would you conclude from these 2 difficulty indices that the item is really more difficult for 12-year-olds? State the significance level associated with your conclusion.
66. If a political issue is favored by 55 per cent of a sample of 200 Republicans, and by 46 per cent of a sample of 250 Democrats, would you con-

Table IV. DATA FOR PASSING (P) AND FAILING (F) ITEMS ON THE STANFORD-BINET TEST

4-year-olds						5-year-olds					
Item			Item			Item			Item		
Case	a	b	Case	a	b	Case	a	b	Case	a	b
1	F	F	21	P	F	41	P	P	61	P	P
2	P	F	22	P	F	42	P	P	62	P	P
3	P	P	23	P	F	43	P	F	63	P	F
4	F	P	24	P	P	44	F	F	64	P	P
5	P	F	25	P	F	45	P	P	65	F	F
6	F	F	26	P	P	46	P	P	66	P	P
7	F	F	27	P	F	47	F	F	67	F	F
8	P	F	28	F	F	48	P	P	68	P	P
9	P	F	29	P	P	49	P	P	69	P	P
10	F	F	30	P	F	50	P	P	70	F	F
11	P	P	31	P	P	51	P	P	71	P	P
12	P	P	32	P	F	52	P	P	72	F	F
13	F	F	33	P	F	53	P	P	73	P	F
14	P	P	34	F	F	54	P	P	74	P	F
15	P	F	35	P	P	55	P	F	75	F	F
16	P	F	36	P	F	56	P	F	76	F	F
17	P	P	37	P	P	57	P	P	77	P	F
18	F	F	38	F	F	58	P	P	78	P	F
19	F	F	39	F	F	59	P	F	79	F	P
20	P	P	40	P	F	60	P	F	80	P	F

clude that the populations of Republicans and Democrats differ on the issue?

67. *a.* Given the data in Table IV, do items *a* and *b* differ significantly in difficulty for the 4-year-olds? Ditto, the 5-year-olds?  
*b.* Is there a significant difference between 4- and 5-year-olds on item *a*? On item *b*?
68. *a.* Would you conclude from the data of Table V that items *c* and *d* differ significantly in difficulty for the 6-year-olds? Ditto, the 7-year-olds?  
*b.* Would you conclude from the data of Table V that, in general, 7-year-olds are more successful than 6-year-olds on item *c*? On item *d*?

Table V. PASSING (P) AND FAILING (F) INFORMATION ON TWO BINET TEST ITEMS AT TWO AGE LEVELS

6-year-olds						7-year-olds					
Item			Item			Item			Item		
Case	<i>c</i>	<i>d</i>	Case	<i>c</i>	<i>d</i>	Case	<i>c</i>	<i>d</i>	Case	<i>c</i>	<i>d</i>
1	F	F	21	P	P	41	P	F	61	P	F
2	P	P	22	P	F	42	P	P	62	F	F
3	F	P	23	F	F	43	F	P	63	P	F
4	F	F	24	F	F	44	P	P	64	F	F
5	F	F	25	F	F	45	F	P	65	P	P
6	F	F	26	P	P	46	F	F	66	P	P
7	P	F	27	P	F	47	P	P	67	P	P
8	F	F	28	P	F	48	F	P	68	P	P
9	P	F	29	F	F	49	P	P	69	P	P
10	F	F	30	F	F	50	P	F	70	P	P
11	P	F	31	F	F	51	F	F	71	P	P
12	P	P	32	F	F	52	F	F	72	P	P
13	F	F	33	P	F	53	F	F	73	P	F
14	P	F	34	P	F	54	P	P	74	P	F
15	F	F	35	F	F	55	F	F	75	F	F
16	P	P	36	F	F	56	P	P	76	P	P
17	F	F	37	F	P	57	F	F	77	P	F
18	F	F	38	P	F	58	P	P	78	P	F
19	F	F	39	F	F	59	P	P	79	P	F
20	F	F	40	P	P	60	F	F	80	P	P

## Chapter 6

69. If you tossed 5 coins 100 times and obtained 2.8 as the mean number of heads, would you suspect bias in the coins? Why? (Be specific.)
70. If you tossed 4 coins 100 times and obtained 2.4 as the mean number of heads, would you suspect bias in the coins? Why? (Be specific.)

71. For a sample of 2970 cases, ages 2.5 to 18, the distribution of IQ's on Form L of the 1937 Stanford-Binet yields:

Mean = 104.00  
SD = 17.03

Skewness ( $g_1$ ) = .028  
Kurtosis ( $g_2$ ) = .346

In answering the following questions, indicate the steps in your computations.

- Would you conclude that the mean IQ of the population for these ages is 100 (the value expected for a properly constructed IQ test)?
  - Is it reasonable to believe that the IQ distribution for the population, at these ages, has normal skewness?
  - Would you conclude from the sample kurtosis that the kurtosis for the population differs from normal kurtosis?
72. Suppose that the mean IQ for the general population is 100 and the standard deviation is 17. If a sample of 289 cases were drawn at random, what would be the probability of obtaining a mean as great as 101? As low as 98?
73. Suppose it is known that the standard deviation of scores for a population is 20. How many cases would you need to draw in order that the standard error of
- a sample mean be 2 score points?
  - a sample SD be 3 points?
74. Suppose that you are polling on an issue for which opinion seems about equally divided. How many cases (how large an  $N$ ) would you need to be sure (at the .01 level of significance) that a sample deviation of 3 per cent from 50 per cent is nonchance?
75. One of the requirements of a good IQ test is that the mean IQ for unselected cases of any school age group shall be 100, and that the distributions for the several age groups shall have the same standard deviations. Given the following for the 1937 Stanford-Binet Test:

Age	6	12
$N$	203	202
$M$	101.0	103.6
$SD$	12.5	20.0

- Is it reasonable to believe that the test is yielding the desired mean when used with 12-year-olds?
  - Would you judge from the results for these 2 age groups that the requirement of equal variability has been met?
76. The means and standard deviations for 2 groups of twins on spool packing are as follows:

	Fraternals	Identicals
$N$	92	94
$M$	761	741
$SD$	79	66

Do these groups differ significantly in mean performance? In variability?

77. Two forms of a test, to be comparable, should yield similar means and similar standard deviations when given to a group. For 202 cases of age 7, we have the following data for the 1937 Stanford-Binet.

	Form L	Form M
$M$	101.8	103.5
$SD$	16.2	15.6

In order to balance practice effect, one-half the group was tested on Form L, then on Form M, while the reverse order was used for the other half. The correlation between the 2 sets of IQ's was .93. Is the obtained difference between means larger than one would expect on the basis of chance sampling? Ditto, the difference between the  $SD$ 's?

78. Measurements on 1000 of each sex at birth have been reported in the literature. The mean length of boys (in centimeters) was 50.51 and the  $SD$  was 2.99, and the values for the girls were 49.90 and 3.00. Is there evidence here for sex difference in length at birth?
79. a. Given that the data for Group A in Table VI were collected to evaluate the effect of practice from a first to a second administration of the same test. Is there evidence for a significant increase in performance?

Table VI. FIRST ( $X_1$ ) AND SECOND ( $X_2$ ) TEST SCORES FOR TWO GROUPS

Group A							
$X_1$	$X_2$	$X_1$	$X_2$	$X_1$	$X_2$	$X_1$	$X_2$
32	31	43	40	26	24	35	36
34	37	37	41	29	32	42	41
16	20	52	57	40	44	34	36
33	33	43	45	38	40	28	29
30	32	31	36	45	46	36	34
35	35	41	44	29	29	27	29
20	19	36	37	50	52	48	48
27	31	39	39	23	29	35	41
34	34	27	20	37	38	28	31
Group B							
$X_1$	$X_2$	$X_1$	$X_2$	$X_1$	$X_2$	$X_1$	$X_2$
32	31	53	55	25	32	31	34
41	45	34	37	30	30	21	27
19	29	30	35	41	44	39	41
46	50	28	33	37	33	37	38
40	37	44	42	36	34	47	57
24	35	33	35	33	39	29	34
33	41	35	40	51	57	42	50
37	39	38	42	20	20	36	39



- b. Given that between the 2 testings of Group *B*, Table VI, special coaching was provided on how to take the test. Would you judge that the coaching of Group *B* led to an increase in scores which is significantly larger than one would expect on the basis of the practice effect demonstrated in the data for Group *A*?
80. Given that the experimental group of Table VII was provided with an experience which should have produced a shift in scores, whereas the control group was exposed only to those ordinary experiences which were presumably alike for both groups. The experimentals and controls were paired on the basis of the pretest scores—case 1 with case 101, case 2 with 102, . . . case 36 with 136. Did the provided experience have a demonstrable effect?

Table VII. BEFORE ( $X_1$ ) AND AFTER ( $X_2$ ) MEASURES ON TWO GROUPS MATCHED BY PAIRING

Experimental Group											
Case	$X_1$	$X_2$	Case	$X_1$	$X_2$	Case	$X_1$	$X_2$	Case	$X_1$	$X_2$
1	66	76	10	75	81	19	62	64	28	64	73
2	78	88	11	57	64	20	78	95	29	71	70
3	62	63	12	68	69	21	60	60	30	58	63
4	52	56	13	51	50	22	54	53	31	70	76
5	37	34	14	61	77	23	59	62	32	75	84
6	65	69	15	70	85	24	66	81	33	70	75
7	60	66	16	76	86	25	72	83	34	67	78
8	85	93	17	63	64	26	86	96	35	57	64
9	64	66	18	46	49	27	49	46	36	55	63
Control Group											
Case	$X_1$	$X_2$	Case	$X_1$	$X_2$	Case	$X_1$	$X_2$	Case	$X_1$	$X_2$
101	67	71	110	74	77	119	61	63	128	64	67
102	77	84	111	56	57	120	80	88	129	71	67
103	62	64	112	68	66	121	61	63	130	58	59
104	53	51	113	50	43	122	54	53	131	70	76
105	41	37	114	61	72	123	58	59	132	75	85
106	65	68	115	70	74	124	66	79	133	68	72
107	59	61	116	76	82	125	73	79	134	67	75
108	83	92	117	63	66	126	90	98	135	57	60
109	65	57	118	45	49	127	47	44	136	55	55

### Chapters 8 and 9

81. a. Using the data of Table VIII, make a scatter diagram with "Ex" on the  $y$  axis, intervals of size 5; and with "TMT" on the  $x$  axis, with  $i = 3$  and the first interval taken as 105-107 (interval sizes are suggested in order to facilitate an exact check of the tallying and subsequent computations).
- b. From the scatter diagram, compute the correlation between "Ex" and "TMT"; also compute the 2 means and the 2 standard deviations.
- c. Write the regression equation for predicting "Ex" from "TMT." Draw the regression line on your scatter diagram.
- d. Determine the error of estimate for predicting "Ex" from a knowledge of "TMT."
- e. What percentage of the variance in "Ex" is due to or associated with variation in "TMT"?
82. Do exercise (81) with "CM" substituted for "TMT" (an appropriate interval size for "CM" is rather obvious).

*Table VIII. DATA FOR 38 STUDENTS IN A COURSE ON MENTAL TESTS*

("Ex" stands for final examination scores; "TMT" stands for IQ's based on Terman-McNemar Test of Mental Ability; "CM" stands for scores on the Terman Concept Mastery Test.)

Ex	TMT	CM	Ex	TMT	CM	Ex	TMT	CM
62	123	47	106	125	126	54	128	69
107	129	59	79	109	33	86	132	82
87	131	78	84	120	56	92	114	41
95	129	74	100	129	81	67	113	44
100	122	52	78	112	51	102	141	112
87	136	127	90	132	110	79	132	72
87	125	74	85	126	54	82	126	54
64	121	46	58	111	33	96	131	111
89	131	97	110	138	119	77	131	65
58	128	71	115	131	138	75	109	22
84	123	28	68	129	39	93	131	108
80	127	53	78	123	67	67	106	25
82	120	53	80	136	101			

83. Using the data of Table IX, compute a "reliability coefficient" for Stanford-Binet IQ's, and determine the standard error of measurement. (Although the data in Table IX are in general typical of those used in determining test reliability by the form vs. form method, they are definitely atypical as regards the Stanford-Binet Test. The inquisitive student may wish to read Chapter 6 in Q. McNemar's *The revision of the Stanford-Binet Scale*, Boston: Houghton Mifflin, 1942.)

Table IX. IQ SCORES FOR 100 CHILDREN WHO WERE TESTED ON BOTH  
FORM L AND FORM M OF THE 1937 STANFORD-BINET TEST

L	M	L	M	L	M	L	M
89	85	93	86	95	93	88	99
88	94	70	74	63	65	88	89
90	89	123	121	104	108	77	71
98	96	93	86	124	125	99	98
110	107	88	105	89	89	77	82
118	121	76	72	110	98	94	98
117	118	98	100	110	110	96	93
113	110	116	115	122	118	70	70
119	119	108	112	126	125	106	106
102	110	113	119	113	122	87	91
99	93	125	121	99	102	122	115
100	104	96	89	104	93	63	61
89	81	135	134	120	118	113	105
98	100	83	92	111	112	64	64
118	112	112	115	91	88	96	94
95	103	92	91	116	114	107	100
133	133	77	82	85	87	98	91
93	93	74	66	86	83	94	93
116	102	83	76	113	107	87	96
143	147	85	83	99	99	104	110
94	98	94	90	103	102	82	78
93	96	86	82	127	126	100	102
124	130	123	124	110	117	78	78
124	119	107	112	138	140	97	95
129	127	138	124	93	96	107	110

APPENDIX  
**Tables A to G**

# Appendix

Table A. NORMAL CURVE FUNCTIONS

$z$ or $x/\sigma$	Area: $m$ to $z$	Area: $q$ Smaller	$y$ or Ordinate
.00	.00000	.50000	.3989
.05	.01994	.48006	.3984
.10	.03983	.46017	.3970
.15	.05962	.44038	.3945
.20	.07926	.42074	.3910
.25	.09871	.40129	.3867
.30	.11791	.38209	.3814
.35	.13683	.36317	.3752
.40	.15542	.34458	.3683
.45	.17364	.32636	.3605
.50	.19146	.30854	.3521
.55	.20884	.29116	.3429
.60	.22575	.27425	.3332
.65	.24215	.25758	.3230
.70	.25804	.24196	.3123
.75	.27337	.22663	.3011
.80	.28814	.21186	.2897
.85	.30234	.19766	.2780
.90	.31594	.18406	.2661
.95	.32894	.17106	.2541
1.00	.34134	.15866	.2420
1.05	.35314	.14686	.2299
1.10	.36433	.13567	.2179
1.15	.37493	.12507	.2059
1.20	.38493	.11507	.1942
1.25	.39435	.10565	.1826
1.30	.40320	.09680	.1714
1.35	.41149	.08851	.1604
1.40	.41924	.08076	.1497
1.45	.42647	.07353	.1394

Table A. NORMAL CURVE FUNCTIONS (*Continued*)

$z$ or $x/\sigma$	Area: $m$ to $z$	Area: $q$ Smaller	$y$ or Ordinate
1.50	.43319	.06681	.1295
1.55	.43943	.06057	.1200
1.60	.44520	.05480	.1109
1.65	.45053	.04947	.1023
1.70	.45543	.04457	.0940
1.75	.45994	.04006	.0863
1.80	.46407	.03593	.0790
1.85	.46784	.03216	.0721
1.90	.47128	.02872	.0656
1.95	.47441	.02559	.0596
2.00	.47725	.02275	.0540
2.05	.47982	.02018	.0488
2.10	.48214	.01786	.0440
2.15	.48422	.01578	.0396
2.20	.48610	.01390	.0355
2.25	.48778	.01222	.0317
2.30	.48928	.01072	.0283
2.35	.49061	.00939	.0252
2.40	.49180	.00820	.0224
2.45	.49286	.00714	.0198
2.50	.49379	.00621	.0175
2.55	.49461	.00539	.0154
2.60	.49534	.00466	.0136
2.65	.49598	.00402	.0119
2.70	.49653	.00347	.0104
2.75	.49702	.00298	.0091
2.80	.49744	.00256	.0079
2.85	.49781	.00219	.0069
2.90	.49813	.00187	.0060
2.95	.49841	.00159	.0051
3.00	.49865	.00135	.0044
3.25	.49942	.00058	.0020
3.50	.49977	.00023	.0009
3.75	.49991	.00009	.0004
4.00	.49997	.00003	.0001



Table B. TRANSFORMATION OF  $r$  TO  $z$ 

$r$	$z$	$r$	$z$	$r$	$z$
.01	.010	.34	.354	.67	.811
.02	.020	.35	.366	.68	.829
.03	.030	.36	.377	.69	.848
.04	.040	.37	.389	.70	.867
.05	.050	.38	.400	.71	.887
.06	.060	.39	.412	.72	.908
.07	.070	.40	.424	.73	.929
.08	.080	.41	.436	.74	.950
.09	.090	.42	.448	.75	.973
.10	.100	.43	.460	.76	.996
.11	.110	.44	.472	.77	1.020
.12	.121	.45	.485	.78	1.045
.13	.131	.46	.497	.79	1.071
.14	.141	.47	.510	.80	1.099
.15	.151	.48	.523	.81	1.127
.16	.161	.49	.536	.82	1.157
.17	.172	.50	.549	.83	1.188
.18	.181	.51	.563	.84	1.221
.19	.192	.52	.577	.85	1.256
.20	.203	.53	.590	.86	1.293
.21	.214	.54	.604	.87	1.333
.22	.224	.55	.618	.88	1.376
.23	.234	.56	.633	.89	1.422
.24	.245	.57	.648	.90	1.472
.25	.256	.58	.663	.91	1.528
.26	.266	.59	.678	.92	1.589
.27	.277	.60	.693	.93	1.658
.28	.288	.61	.709	.94	1.738
.29	.299	.62	.725	.95	1.832
.30	.309	.63	.741	.96	1.946
.31	.321	.64	.758	.97	2.092
.32	.332	.65	.775	.98	2.298
.33	.343	.66	.793	.99	2.647

Table C. TRANSFORMATION OF  $z$  TO  $r$  \*

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.0000	.0100	.0200	.0300	.0400	.0500	.0599	.0699	.0798	.0898
.1	.0997	.1096	.1194	.1293	.1391	.1489	.1586	.1684	.1781	.1877
.2	.1974	.2070	.2165	.2260	.2355	.2449	.2543	.2636	.2729	.2821
.3	.2913	.3004	.3095	.3185	.3275	.3364	.3452	.3540	.3627	.3714
.4	.3800	.3885	.3969	.4053	.4136	.4219	.4301	.4382	.4462	.4542
.5	.4621	.4699	.4777	.4854	.4930	.5005	.5080	.5154	.5227	.5299
.6	.5370	.5441	.5511	.5580	.5649	.5717	.5784	.5850	.5915	.5980
.7	.6044	.6107	.6169	.6231	.6291	.6351	.6411	.6469	.6527	.6584
.8	.6640	.6696	.6751	.6805	.6858	.6911	.6963	.7014	.7064	.7114
.9	.7163	.7211	.7259	.7306	.7352	.7398	.7443	.7487	.7531	.7574
1.0	.7616	.7658	.7699	.7739	.7779	.7818	.7857	.7895	.7932	.7969
1.1	.8005	.8041	.8076	.8110	.8144	.8178	.8210	.8243	.8275	.8306
1.2	.8337	.8367	.8397	.8426	.8455	.8483	.8511	.8538	.8565	.8591
1.3	.8617	.8643	.8668	.8692	.8717	.8741	.8764	.8787	.8810	.8832
1.4	.8854	.8875	.8896	.8917	.8937	.8957	.8977	.8996	.9015	.9033
1.5	.9051	.9069	.9087	.9104	.9121	.9138	.9154	.9170	.9186	.9201
1.6	.9217	.9232	.9246	.9261	.9275	.9289	.9302	.9316	.9329	.9341
1.7	.9354	.9366	.9379	.9391	.9402	.9414	.9425	.9436	.9447	.9458
1.8	.9468	.9478	.9488	.9498	.9508	.9518	.9527	.9536	.9545	.9554
1.9	.9562	.9571	.9579	.9587	.9595	.9603	.9611	.9618	.9626	.9633
2.0	.9640	.9647	.9654	.9661	.9668	.9674	.9680	.9686	.9693	.9699
2.1	.9704	.9710	.9716	.9722	.9727	.9732	.9738	.9743	.9748	.9753
2.2	.9757	.9762	.9767	.9771	.9776	.9780	.9785	.9789	.9793	.9797
2.3	.9801	.9805	.9809	.9812	.9816	.9820	.9823	.9827	.9830	.9834
2.4	.9837	.9840	.9843	.9846	.9849	.9852	.9855	.9858	.9861	.9864
2.5	.9866	.9869	.9871	.9874	.9876	.9879	.9881	.9884	.9886	.9888
2.6	.9890	.9892	.9894	.9897	.9899	.9901	.9903	.9904	.9906	.9908
2.7	.9910	.9912	.9914	.9915	.9917	.9919	.9920	.9922	.9923	.9925
2.8	.9926	.9928	.9929	.9931	.9932	.9933	.9935	.9936	.9937	.9938
2.9	.9940	.9941	.9942	.9943	.9944	.9945	.9946	.9948	.9948	.9950

\* Table C is abridged from Table VII of Fisher and Yates: *Statistical tables for biological, agricultural and medical research*, Oliver and Boyd, Ltd., Edinburgh, by permission of the authors and publishers.

Table D. DISTRIBUTION OF  $\chi^2$ \*

$n$	$P = .99$	.98	.95	.90	.80	.70	.50
1	.00016	.00063	.0039	.016	.064	.15	.46
2	.02	.04	.10	.21	.45	.71	1.89
3	.12	.18	.35	.58	1.00	1.42	2.37
4	.30	.43	.71	1.06	1.65	2.20	3.36
5	.55	.75	1.14	1.61	2.34	3.00	4.35
6	.87	1.13	1.64	2.20	3.07	3.83	5.35
7	1.24	1.56	2.17	2.83	3.82	4.67	6.35
8	1.65	2.03	2.73	3.49	4.59	5.53	7.34
9	2.09	2.53	3.32	4.17	5.38	6.39	8.34
10	2.56	3.06	3.94	4.86	6.18	7.27	9.34
11	3.05	3.61	4.58	5.58	6.99	8.15	10.34
12	3.57	4.18	5.23	6.30	7.81	9.03	11.34
13	4.11	4.76	5.89	7.04	8.63	9.93	12.34
14	4.66	5.37	6.57	7.79	9.47	10.82	13.34
15	5.23	5.98	7.26	8.55	10.31	11.72	14.34
16	5.81	6.61	7.96	9.31	11.15	12.62	15.34
17	6.41	7.26	8.67	10.08	12.00	13.53	16.34
18	7.02	7.91	9.39	10.86	12.86	14.44	17.34
19	7.63	8.57	10.12	11.65	13.72	15.35	18.34
20	8.26	9.24	10.85	12.44	14.58	16.27	19.34
21	8.90	9.92	11.59	13.24	15.44	17.18	20.34
22	9.54	10.60	12.34	14.04	16.31	18.10	21.34
23	10.20	11.29	13.09	14.85	17.19	19.02	22.34
24	10.86	11.99	13.85	15.66	18.06	19.94	23.34
25	11.52	12.70	14.61	16.47	18.94	20.87	24.34
26	12.20	13.41	15.38	17.29	19.82	21.79	25.34
27	12.88	14.12	16.15	18.11	20.70	22.72	26.34
28	13.56	14.85	16.93	18.94	21.59	23.65	27.34
29	14.26	15.57	17.71	19.77	22.48	24.58	28.34
30	14.95	16.31	18.49	20.60	23.36	25.51	29.34

\* Table D is abridged from Table IV of Fisher and Yates: *Statistical tables for biological, agricultural and medical research*, Oliver and Boyd, Ltd., Edinburgh, by permission of the authors and publishers.

Table D. DISTRIBUTION OF  $\chi^2$ \*—(Continued)

$n$	.30	.20	.10	.05	.02	.01	.001
1	1.07	1.64	2.71	3.84	5.41	6.64	10.83
2	2.41	3.22	4.60	5.99	7.82	9.21	13.82
3	3.66	4.64	6.25	7.82	9.84	11.34	16.27
4	4.88	5.99	7.78	9.49	11.67	13.28	18.46
5	6.06	7.29	9.24	11.07	13.39	15.09	20.52
6	7.23	8.56	10.64	12.59	15.03	16.81	22.46
7	8.38	9.80	12.02	14.07	16.62	18.48	24.32
8	9.52	11.03	13.36	15.51	18.17	20.09	26.12
9	10.66	12.24	14.68	16.92	19.68	21.67	27.88
10	11.78	13.44	15.99	18.31	21.16	23.21	29.59
11	12.90	14.63	17.28	19.68	22.62	24.72	31.26
12	14.01	15.81	18.55	21.03	24.05	26.22	32.91
13	15.12	16.98	19.81	22.36	25.47	27.69	34.53
14	16.22	18.15	21.06	23.68	26.87	29.14	36.12
15	17.32	19.31	22.31	25.00	28.26	30.58	37.70
16	18.42	20.46	23.54	26.30	29.63	32.00	39.25
17	19.51	21.62	24.77	27.59	31.00	33.41	40.79
18	20.60	22.76	25.99	28.87	32.35	34.80	42.31
19	21.69	23.90	27.20	30.14	33.69	36.19	43.82
20	22.78	25.04	28.41	31.41	35.02	37.57	45.32
21	23.86	26.17	29.62	32.67	36.34	38.93	46.80
22	24.94	27.30	30.81	33.92	37.66	40.29	48.27
23	26.02	28.43	32.01	35.17	38.97	41.64	49.73
24	27.10	29.55	33.20	36.42	40.27	42.98	51.18
25	28.17	30.68	34.38	37.65	41.57	44.31	52.62
26	29.25	31.80	35.56	38.88	42.86	45.64	54.05
27	30.32	32.91	36.74	40.11	44.14	46.96	55.48
28	31.39	34.03	37.92	41.34	45.42	48.28	56.89
29	32.46	35.14	39.09	42.56	46.69	49.59	58.30
30	33.53	36.25	40.26	43.77	47.96	50.89	59.70

\* Table D is abridged from Table IV of Fisher and Yates: *Statistical tables for biological, agricultural and medical research*, Oliver and Boyd, Ltd., Edinburgh, by permission of the authors and publishers

Table E. DISTRIBUTION OF  $t$  \*

$n$	$P = .1$	.05	.02	.01	.001
1	6.314	12.706	31.821	63.657	636.619
2	2.920	4.303	6.965	9.925	31.598
3	2.353	3.182	4.541	5.841	12.941
4	2.132	2.776	3.747	4.604	8.610
5	2.015	2.571	3.365	4.032	6.859
6	1.943	2.447	3.143	3.707	5.959
7	1.895	2.365	2.998	3.499	5.405
8	1.860	2.306	2.896	3.355	5.041
9	1.833	2.262	2.821	3.250	4.781
10	1.812	2.228	2.764	3.169	4.587
11	1.796	2.201	2.718	3.106	4.437
12	1.782	2.179	2.681	3.055	4.318
13	1.771	2.160	2.650	3.012	4.221
14	1.761	2.145	2.624	2.977	4.140
15	1.753	2.131	2.602	2.947	4.073
16	1.746	2.120	2.583	2.921	4.015
17	1.740	2.110	2.567	2.898	3.965
18	1.734	2.101	2.552	2.878	3.922
19	1.729	2.093	2.539	2.861	3.883
20	1.725	2.086	2.528	2.845	3.850
21	1.721	2.080	2.518	2.831	3.819
22	1.717	2.074	2.508	2.819	3.792
23	1.714	2.069	2.500	2.807	3.767
24	1.711	2.064	2.492	2.797	3.745
25	1.708	2.060	2.485	2.787	3.725
26	1.706	2.056	2.479	2.779	3.707
27	1.703	2.052	2.473	2.771	3.690
28	1.701	2.048	2.467	2.763	3.674
29	1.699	2.045	2.462	2.756	3.659
30	1.697	2.042	2.457	2.750	3.646
40	1.684	2.021	2.423	2.704	3.551
60	1.671	2.000	2.390	2.660	3.460
120	1.658	1.980	2.358	2.617	3.373
$\infty$	1.645	1.960	2.326	2.576	3.291

\* Table E is abridged from Table III of Fisher and Yates: *Statistical tables for biological, agricultural and medical research*, Oliver and Boyd, Ltd., Edinburgh, by permission of the authors and publishers.

Table F. TABLE OF *F* FOR .05 (roman), .01 (*italic*), AND .001 (bold face) LEVELS OF SIGNIFICANCE \*

$\frac{n_1}{n_2}$	1	2	3	4	5	6	8	12	24	$\infty$
1	161	200	216	225	230	234	239	244	249	254
	<i>4052</i>	<i>4999</i>	<i>5403</i>	<i>5625</i>	<i>5724</i>	<i>5859</i>	<i>5981</i>	<i>6106</i>	<i>6234</i>	<i>6366</i>
	<b>405284</b>	<b>500000</b>	<b>540379</b>	<b>562500</b>	<b>576405</b>	<b>585937</b>	<b>598144</b>	<b>610667</b>	<b>623497</b>	<b>636619</b>
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
	<i>98.49</i>	<i>99.01</i>	<i>99.17</i>	<i>99.25</i>	<i>99.30</i>	<i>99.33</i>	<i>99.36</i>	<i>99.40</i>	<i>99.43</i>	<i>99.50</i>
	<b>998.5</b>	<b>999.0</b>	<b>999.2</b>	<b>999.2</b>	<b>999.3</b>	<b>999.3</b>	<b>999.4</b>	<b>999.4</b>	<b>999.5</b>	<b>999.5</b>
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
	<i>34.12</i>	<i>30.81</i>	<i>29.46</i>	<i>28.71</i>	<i>28.24</i>	<i>27.91</i>	<i>27.49</i>	<i>27.05</i>	<i>26.60</i>	<i>26.12</i>
	<b>167.5</b>	<b>148.5</b>	<b>141.1</b>	<b>137.1</b>	<b>134.6</b>	<b>132.8</b>	<b>130.6</b>	<b>128.3</b>	<b>125.9</b>	<b>123.5</b>
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
	<i>21.20</i>	<i>18.00</i>	<i>16.69</i>	<i>15.98</i>	<i>15.59</i>	<i>15.21</i>	<i>14.80</i>	<i>14.37</i>	<i>13.93</i>	<i>13.46</i>
	<b>74.14</b>	<b>61.36</b>	<b>56.18</b>	<b>53.44</b>	<b>51.71</b>	<b>50.53</b>	<b>49.00</b>	<b>47.41</b>	<b>45.77</b>	<b>44.06</b>
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
	<i>16.26</i>	<i>13.27</i>	<i>12.06</i>	<i>11.39</i>	<i>10.97</i>	<i>10.67</i>	<i>10.27</i>	<i>9.89</i>	<i>9.47</i>	<i>9.02</i>
	<b>47.04</b>	<b>36.61</b>	<b>33.20</b>	<b>31.09</b>	<b>29.75</b>	<b>28.84</b>	<b>27.64</b>	<b>26.42</b>	<b>25.14</b>	<b>23.78</b>
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
	<i>13.74</i>	<i>10.92</i>	<i>9.78</i>	<i>9.15</i>	<i>8.75</i>	<i>8.47</i>	<i>8.10</i>	<i>7.72</i>	<i>7.31</i>	<i>6.88</i>
	<b>35.51</b>	<b>27.00</b>	<b>23.70</b>	<b>21.90</b>	<b>20.81</b>	<b>20.03</b>	<b>19.03</b>	<b>17.99</b>	<b>16.89</b>	<b>15.75</b>
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
	<i>12.25</i>	<i>9.55</i>	<i>8.45</i>	<i>7.85</i>	<i>7.46</i>	<i>7.19</i>	<i>6.84</i>	<i>6.47</i>	<i>6.07</i>	<i>5.65</i>
	<b>29.22</b>	<b>21.69</b>	<b>18.77</b>	<b>17.19</b>	<b>16.21</b>	<b>15.52</b>	<b>14.63</b>	<b>13.71</b>	<b>12.73</b>	<b>11.69</b>
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
	<i>11.26</i>	<i>8.65</i>	<i>7.59</i>	<i>7.01</i>	<i>6.63</i>	<i>6.37</i>	<i>6.03</i>	<i>5.67</i>	<i>5.28</i>	<i>4.89</i>
	<b>25.42</b>	<b>18.49</b>	<b>15.83</b>	<b>14.39</b>	<b>13.49</b>	<b>12.86</b>	<b>12.04</b>	<b>11.19</b>	<b>10.30</b>	<b>9.34</b>
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
	<i>10.56</i>	<i>8.02</i>	<i>6.99</i>	<i>6.42</i>	<i>6.06</i>	<i>5.80</i>	<i>5.47</i>	<i>5.11</i>	<i>4.73</i>	<i>4.31</i>
	<b>22.36</b>	<b>16.39</b>	<b>13.90</b>	<b>12.56</b>	<b>11.71</b>	<b>11.13</b>	<b>10.37</b>	<b>9.57</b>	<b>8.72</b>	<b>7.81</b>
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
	<i>10.04</i>	<i>7.56</i>	<i>6.55</i>	<i>5.99</i>	<i>5.64</i>	<i>5.39</i>	<i>5.06</i>	<i>4.71</i>	<i>4.33</i>	<i>3.91</i>
	<b>21.04</b>	<b>14.91</b>	<b>12.55</b>	<b>11.28</b>	<b>10.48</b>	<b>9.92</b>	<b>9.20</b>	<b>8.45</b>	<b>7.64</b>	<b>6.76</b>
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
	<i>9.65</i>	<i>7.20</i>	<i>6.22</i>	<i>5.67</i>	<i>5.32</i>	<i>5.07</i>	<i>4.74</i>	<i>4.40</i>	<i>4.02</i>	<i>3.60</i>
	<b>19.69</b>	<b>13.81</b>	<b>11.56</b>	<b>10.35</b>	<b>9.58</b>	<b>9.05</b>	<b>8.35</b>	<b>7.63</b>	<b>6.85</b>	<b>6.00</b>
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
	<i>9.33</i>	<i>6.93</i>	<i>5.95</i>	<i>5.41</i>	<i>5.06</i>	<i>4.82</i>	<i>4.50</i>	<i>4.16</i>	<i>3.78</i>	<i>3.36</i>
	<b>18.64</b>	<b>12.97</b>	<b>10.80</b>	<b>9.63</b>	<b>8.89</b>	<b>8.38</b>	<b>7.71</b>	<b>7.00</b>	<b>6.25</b>	<b>5.42</b>

\* Table F is reprinted, in rearranged form, from Table V of Fisher and Yates: *Statistical tables for biological, agricultural and medical research*, Oliver and Boyd, Ltd., Edinburgh, by permission of the authors and publishers.



Table F. TABLE OF *F* FOR .05 (roman), .01 (*italic*), AND .001 (bold face)  
LEVELS OF SIGNIFICANCE \*—(Continued)

$n_1$ $n_2$	1	2	3	4	5	6	8	12	24	$\infty$
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
	<i>9.07</i>	<i>6.70</i>	<i>5.74</i>	<i>5.20</i>	<i>4.88</i>	<i>4.62</i>	<i>4.30</i>	<i>3.96</i>	<i>3.59</i>	<i>3.16</i>
	<b>17.81</b>	<b>12.31</b>	<b>10.21</b>	<b>9.07</b>	<b>8.35</b>	<b>7.86</b>	<b>7.21</b>	<b>6.52</b>	<b>5.78</b>	<b>4.97</b>
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
	<i>8.86</i>	<i>6.51</i>	<i>5.56</i>	<i>5.03</i>	<i>4.69</i>	<i>4.46</i>	<i>4.14</i>	<i>3.80</i>	<i>3.43</i>	<i>3.00</i>
	<b>17.14</b>	<b>11.78</b>	<b>9.73</b>	<b>8.62</b>	<b>7.92</b>	<b>7.43</b>	<b>6.80</b>	<b>6.13</b>	<b>5.41</b>	<b>4.60</b>
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
	<i>8.68</i>	<i>6.36</i>	<i>5.42</i>	<i>4.89</i>	<i>4.56</i>	<i>4.33</i>	<i>4.00</i>	<i>3.67</i>	<i>3.29</i>	<i>2.87</i>
	<b>16.59</b>	<b>11.34</b>	<b>9.34</b>	<b>8.25</b>	<b>7.57</b>	<b>7.09</b>	<b>6.47</b>	<b>5.81</b>	<b>5.10</b>	<b>4.31</b>
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
	<i>8.53</i>	<i>6.23</i>	<i>5.29</i>	<i>4.77</i>	<i>4.44</i>	<i>4.20</i>	<i>3.89</i>	<i>3.55</i>	<i>3.18</i>	<i>2.76</i>
	<b>16.12</b>	<b>10.97</b>	<b>9.00</b>	<b>7.94</b>	<b>7.27</b>	<b>6.81</b>	<b>6.19</b>	<b>5.55</b>	<b>4.85</b>	<b>4.06</b>
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
	<i>8.40</i>	<i>6.11</i>	<i>5.18</i>	<i>4.67</i>	<i>4.34</i>	<i>4.10</i>	<i>3.79</i>	<i>3.45</i>	<i>3.08</i>	<i>2.66</i>
	<b>15.72</b>	<b>10.66</b>	<b>8.73</b>	<b>7.68</b>	<b>7.02</b>	<b>6.56</b>	<b>5.96</b>	<b>5.32</b>	<b>4.63</b>	<b>3.85</b>
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
	<i>8.28</i>	<i>6.01</i>	<i>5.09</i>	<i>4.58</i>	<i>4.25</i>	<i>4.01</i>	<i>3.71</i>	<i>3.37</i>	<i>3.00</i>	<i>2.57</i>
	<b>15.38</b>	<b>10.39</b>	<b>8.49</b>	<b>7.46</b>	<b>6.81</b>	<b>6.35</b>	<b>5.76</b>	<b>5.13</b>	<b>4.45</b>	<b>3.67</b>
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
	<i>8.18</i>	<i>5.93</i>	<i>5.01</i>	<i>4.60</i>	<i>4.17</i>	<i>3.94</i>	<i>3.63</i>	<i>3.30</i>	<i>2.92</i>	<i>2.49</i>
	<b>15.08</b>	<b>10.16</b>	<b>8.28</b>	<b>7.26</b>	<b>6.61</b>	<b>6.18</b>	<b>5.59</b>	<b>4.97</b>	<b>4.29</b>	<b>3.52</b>
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
	<i>8.10</i>	<i>5.85</i>	<i>4.94</i>	<i>4.43</i>	<i>4.10</i>	<i>3.87</i>	<i>3.56</i>	<i>3.23</i>	<i>2.86</i>	<i>2.42</i>
	<b>14.82</b>	<b>9.96</b>	<b>8.10</b>	<b>7.10</b>	<b>6.46</b>	<b>6.02</b>	<b>5.44</b>	<b>4.82</b>	<b>4.15</b>	<b>3.38</b>
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
	<i>8.02</i>	<i>5.78</i>	<i>4.87</i>	<i>4.37</i>	<i>4.04</i>	<i>3.81</i>	<i>3.51</i>	<i>3.17</i>	<i>2.80</i>	<i>2.36</i>
	<b>14.59</b>	<b>9.77</b>	<b>7.94</b>	<b>6.95</b>	<b>6.32</b>	<b>5.88</b>	<b>5.31</b>	<b>4.70</b>	<b>4.03</b>	<b>3.26</b>
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
	<i>7.94</i>	<i>5.73</i>	<i>4.82</i>	<i>4.31</i>	<i>3.99</i>	<i>3.76</i>	<i>3.45</i>	<i>3.12</i>	<i>2.75</i>	<i>2.31</i>
	<b>14.38</b>	<b>9.61</b>	<b>7.80</b>	<b>6.81</b>	<b>6.19</b>	<b>5.76</b>	<b>5.19</b>	<b>4.58</b>	<b>3.92</b>	<b>3.15</b>
23	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20	2.00	1.76
	<i>7.88</i>	<i>5.66</i>	<i>4.76</i>	<i>4.26</i>	<i>3.94</i>	<i>3.71</i>	<i>3.41</i>	<i>3.07</i>	<i>2.70</i>	<i>2.26</i>
	<b>14.19</b>	<b>9.47</b>	<b>7.67</b>	<b>6.69</b>	<b>6.08</b>	<b>5.65</b>	<b>5.09</b>	<b>4.48</b>	<b>3.82</b>	<b>3.05</b>
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
	<i>7.82</i>	<i>5.61</i>	<i>4.72</i>	<i>4.22</i>	<i>3.90</i>	<i>3.67</i>	<i>3.36</i>	<i>3.03</i>	<i>2.66</i>	<i>2.21</i>
	<b>14.03</b>	<b>9.34</b>	<b>7.55</b>	<b>6.59</b>	<b>5.98</b>	<b>5.55</b>	<b>4.99</b>	<b>4.39</b>	<b>3.74</b>	<b>2.97</b>

\* Table F is reprinted, in rearranged form, from Table V of Fisher and Yates: *Statistical tables for biological, agricultural and medical research*, Oliver and Boyd, Ltd., Edinburgh, by permission of the authors and publishers.

Table F. TABLE OF *F* FOR .05 (roman), .01 (*italic*), AND .001 (bold face)  
LEVELS OF SIGNIFICANCE \*—(Continued)

$n_1 \backslash n_2$	1	2	3	4	5	6	8	12	24	$\infty$
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.98	1.71
	7.77	5.57	4.68	4.18	3.86	3.63	3.38	2.99	2.62	2.17
	<b>13.88</b>	<b>9.22</b>	<b>7.45</b>	<b>6.49</b>	<b>5.88</b>	<b>5.46</b>	<b>4.91</b>	<b>4.31</b>	<b>3.66</b>	<b>2.89</b>
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
	7.82	5.58	4.84	4.14	3.82	3.59	3.29	2.96	2.58	2.13
	<b>13.74</b>	<b>9.12</b>	<b>7.36</b>	<b>6.41</b>	<b>5.80</b>	<b>5.38</b>	<b>4.83</b>	<b>4.24</b>	<b>3.59</b>	<b>2.82</b>
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67
	7.68	5.49	4.60	4.11	3.78	3.56	3.26	2.93	2.55	2.10
	<b>13.61</b>	<b>9.02</b>	<b>7.27</b>	<b>6.33</b>	<b>5.73</b>	<b>5.31</b>	<b>4.76</b>	<b>4.17</b>	<b>3.52</b>	<b>2.75</b>
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
	7.64	5.45	4.57	4.07	3.75	3.53	3.23	2.90	2.52	2.06
	<b>13.50</b>	<b>8.93</b>	<b>7.19</b>	<b>6.25</b>	<b>5.66</b>	<b>5.24</b>	<b>4.69</b>	<b>4.11</b>	<b>3.46</b>	<b>2.70</b>
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
	7.60	5.42	4.54	4.04	3.73	3.50	3.20	2.87	2.49	2.03
	<b>13.39</b>	<b>8.85</b>	<b>7.12</b>	<b>6.19</b>	<b>5.59</b>	<b>5.18</b>	<b>4.64</b>	<b>4.05</b>	<b>3.41</b>	<b>2.64</b>
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.84	2.47	2.01
	<b>13.29</b>	<b>8.77</b>	<b>7.05</b>	<b>6.12</b>	<b>5.53</b>	<b>5.12</b>	<b>4.58</b>	<b>4.00</b>	<b>3.36</b>	<b>2.59</b>
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.66	2.29	1.80
	<b>12.61</b>	<b>8.25</b>	<b>6.60</b>	<b>5.70</b>	<b>5.13</b>	<b>4.73</b>	<b>4.21</b>	<b>3.64</b>	<b>3.01</b>	<b>2.23</b>
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.50	2.12	1.60
	<b>11.97</b>	<b>7.76</b>	<b>6.17</b>	<b>5.31</b>	<b>4.76</b>	<b>4.37</b>	<b>3.87</b>	<b>3.31</b>	<b>2.69</b>	<b>1.90</b>
120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83	1.61	1.25
	6.85	4.79	3.95	3.48	3.17	2.96	2.66	2.34	1.95	1.38
	<b>11.38</b>	<b>7.31</b>	<b>5.79</b>	<b>4.95</b>	<b>4.42</b>	<b>4.04</b>	<b>3.55</b>	<b>3.02</b>	<b>2.40</b>	<b>1.56</b>
$\infty$	3.84	2.99	2.60	2.37	2.21	2.09	1.94	1.75	1.52	1.00
	6.64	4.60	3.78	3.32	3.02	2.80	2.51	2.18	1.79	1.00
	<b>10.83</b>	<b>6.91</b>	<b>5.42</b>	<b>4.62</b>	<b>4.10</b>	<b>3.74</b>	<b>3.27</b>	<b>2.74</b>	<b>2.13</b>	<b>1.00</b>

\* Table F is reprinted, in rearranged form, from Table V of Fisher and Yates: *Statistical tables for biological, agricultural and medical research*, Oliver and Boyd, Ltd., Edinburgh, by permission of the authors and publishers.

Table G. SQUARES AND SQUARE ROOTS

N	N <sup>2</sup>	$\sqrt{N}$	$\sqrt{10N}$
1.00	1.0000	1.00000	3.16228
1.01	1.0201	1.00499	3.17805
1.02	1.0404	1.00995	3.19374
1.03	1.0609	1.01489	3.20936
1.04	1.0816	1.01980	3.22490
1.05	1.1025	1.02470	3.24037
1.06	1.1236	1.02956	3.25576
1.07	1.1449	1.03441	3.27109
1.08	1.1664	1.03925	3.28634
1.09	1.1881	1.04403	3.30151
1.10	1.2100	1.04881	3.31662
1.11	1.2321	1.05357	3.33167
1.12	1.2544	1.05830	3.34664
1.13	1.2769	1.06301	3.36155
1.14	1.2996	1.06771	3.37639
1.15	1.3225	1.07238	3.39116
1.16	1.3456	1.07703	3.40588
1.17	1.3689	1.08167	3.42053
1.18	1.3924	1.08628	3.43511
1.19	1.4161	1.09087	3.44964
1.20	1.4400	1.09545	3.46410
1.21	1.4641	1.10000	3.47851
1.22	1.4884	1.10454	3.49285
1.23	1.5129	1.10905	3.50714
1.24	1.5376	1.11355	3.52136
1.25	1.5625	1.11803	3.53553
1.26	1.5876	1.12250	3.54965
1.27	1.6129	1.12694	3.56371
1.28	1.6384	1.13137	3.57771
1.29	1.6641	1.13578	3.59166
1.30	1.6900	1.14018	3.60555
1.31	1.7161	1.14455	3.61939
1.32	1.7424	1.14891	3.63318
1.33	1.7689	1.15326	3.64692
1.34	1.7956	1.15758	3.66060
1.35	1.8225	1.16190	3.67423
1.36	1.8496	1.16619	3.68782
1.37	1.8769	1.17047	3.70135
1.38	1.9044	1.17473	3.71484
1.39	1.9321	1.17898	3.72827
1.40	1.9600	1.18322	3.74166
1.41	1.9881	1.18743	3.75500
1.42	2.0164	1.19164	3.76829
1.43	2.0449	1.19583	3.78153
1.44	2.0736	1.20000	3.79473
1.45	2.1025	1.20416	3.80789
1.46	2.1316	1.20830	3.82099
1.47	2.1609	1.21244	3.83406
1.48	2.1904	1.21655	3.84708
1.49	2.2201	1.22066	3.86005
1.50	2.2500	1.22474	3.87298
N	N <sup>2</sup>	$\sqrt{N}$	$\sqrt{10N}$

N	N <sup>2</sup>	$\sqrt{N}$	$\sqrt{10N}$
1.50	2.2500	1.22474	3.87298
1.51	2.2801	1.22882	3.88587
1.52	2.3104	1.23288	3.89872
1.53	2.3409	1.23693	3.91152
1.54	2.3716	1.24097	3.92428
1.55	2.4025	1.24499	3.93700
1.56	2.4336	1.24900	3.94968
1.57	2.4649	1.25300	3.96232
1.58	2.4964	1.25698	3.97492
1.59	2.5281	1.26095	3.98748
1.60	2.5600	1.26491	4.00000
1.61	2.5921	1.26886	4.01248
1.62	2.6244	1.27279	4.02492
1.63	2.6569	1.27671	4.03733
1.64	2.6896	1.28062	4.04969
1.65	2.7225	1.28452	4.06202
1.66	2.7556	1.28841	4.07431
1.67	2.7889	1.29228	4.08656
1.68	2.8224	1.29615	4.09878
1.69	2.8561	1.30000	4.11096
1.70	2.8900	1.30384	4.12311
1.71	2.9241	1.30767	4.13521
1.72	2.9584	1.31149	4.14729
1.73	2.9929	1.31529	4.15933
1.74	3.0276	1.31909	4.17133
1.75	3.0625	1.32288	4.18330
1.76	3.0976	1.32665	4.19524
1.77	3.1329	1.33041	4.20714
1.78	3.1684	1.33417	4.21900
1.79	3.2041	1.33791	4.23084
1.80	3.2400	1.34164	4.24264
1.81	3.2761	1.34536	4.25441
1.82	3.3124	1.34907	4.26616
1.83	3.3489	1.35277	4.27785
1.84	3.3856	1.35647	4.28952
1.85	3.4225	1.36015	4.30116
1.86	3.4596	1.36382	4.31277
1.87	3.4969	1.36748	4.32435
1.88	3.5344	1.37113	4.33590
1.89	3.5721	1.37477	4.34741
1.90	3.6100	1.37840	4.35890
1.91	3.6481	1.38203	4.37035
1.92	3.6864	1.38564	4.38178
1.93	3.7249	1.38924	4.39318
1.94	3.7636	1.39284	4.40454
1.95	3.8025	1.39642	4.41588
1.96	3.8416	1.40000	4.42719
1.97	3.8809	1.40357	4.43847
1.98	3.9204	1.40712	4.44972
1.99	3.9601	1.41067	4.46094
2.00	4.0000	1.41421	4.47214
N	N <sup>2</sup>	$\sqrt{N}$	$\sqrt{10N}$

Table G. SQUARES AND SQUARE ROOTS—(Continued)

N	N <sup>2</sup>	√N	√10N
<b>2.00</b>	<b>4.0000</b>	<b>1.41421</b>	<b>4.47214</b>
2.01	4.0401	1.41774	4.48330
2.02	4.0804	1.42127	4.49444
2.03	4.1209	1.42478	4.50555
2.04	4.1616	1.42829	4.51664
2.05	4.2025	1.43178	4.52769
2.06	4.2436	1.43527	4.53872
2.07	4.2849	1.43875	4.54973
2.08	4.3264	1.44222	4.56070
2.09	4.3681	1.44568	4.57165
<b>2.10</b>	<b>4.4100</b>	<b>1.44914</b>	<b>4.58258</b>
2.11	4.4521	1.45258	4.59347
2.12	4.4944	1.45602	4.60435
2.13	4.5369	1.45945	4.61519
2.14	4.5796	1.46287	4.62601
2.15	4.6225	1.46629	4.63681
2.16	4.6656	1.46969	4.64758
2.17	4.7089	1.47309	4.65833
2.18	4.7524	1.47648	4.66905
2.19	4.7961	1.47986	4.67974
<b>2.20</b>	<b>4.8400</b>	<b>1.48324</b>	<b>4.69042</b>
2.21	4.8841	1.48661	4.70106
2.22	4.9284	1.48997	4.71169
2.23	4.9729	1.49332	4.72229
2.24	5.0176	1.49666	4.73286
2.25	5.0625	1.50000	4.74342
2.26	5.1076	1.50333	4.75395
2.27	5.1529	1.50665	4.76445
2.28	5.1984	1.50997	4.77493
2.29	5.2441	1.51327	4.78539
<b>2.30</b>	<b>5.2900</b>	<b>1.51658</b>	<b>4.79583</b>
2.31	5.3361	1.51987	4.80625
2.32	5.3824	1.52315	4.81664
2.33	5.4289	1.52643	4.82701
2.34	5.4756	1.52971	4.83735
2.35	5.5225	1.53297	4.84768
2.36	5.5696	1.53623	4.85798
2.37	5.6169	1.53948	4.86826
2.38	5.6644	1.54272	4.87852
2.39	5.7121	1.54596	4.88876
<b>2.40</b>	<b>5.7600</b>	<b>1.54919</b>	<b>4.89898</b>
2.41	5.8081	1.55242	4.90918
2.42	5.8564	1.55563	4.91935
2.43	5.9049	1.55885	4.92950
2.44	5.9536	1.56205	4.93964
2.45	6.0025	1.56525	4.94975
2.46	6.0516	1.56844	4.95984
2.47	6.1009	1.57162	4.96991
2.48	6.1504	1.57480	4.97996
2.49	6.2001	1.57797	4.98999
<b>2.50</b>	<b>6.2500</b>	<b>1.58114</b>	<b>5.00000</b>
<b>N</b>	<b>N<sup>2</sup></b>	<b>√N</b>	<b>√10N</b>

N	N <sup>2</sup>	√N	√10N
<b>2.50</b>	<b>6.2500</b>	<b>1.58114</b>	<b>5.00000</b>
2.51	6.3001	1.58430	5.00999
2.52	6.3504	1.58745	5.01996
2.53	6.4009	1.59060	5.02991
2.54	6.4516	1.59374	5.03984
2.55	6.5025	1.59687	5.04975
2.56	6.5536	1.60000	5.05964
2.57	6.6049	1.60312	5.06952
2.58	6.6564	1.60624	5.07937
2.59	6.7081	1.60935	5.08920
<b>2.60</b>	<b>6.7600</b>	<b>1.61245</b>	<b>5.09902</b>
2.61	6.8121	1.61555	5.10882
2.62	6.8644	1.61864	5.11859
2.63	6.9169	1.62173	5.12835
2.64	6.9696	1.62481	5.13809
2.65	7.0225	1.62788	5.14782
2.66	7.0756	1.63095	5.15752
2.67	7.1289	1.63401	5.16720
2.68	7.1824	1.63707	5.17687
2.69	7.2361	1.64012	5.18652
<b>2.70</b>	<b>7.2900</b>	<b>1.64317</b>	<b>5.19615</b>
2.71	7.3441	1.64621	5.20577
2.72	7.3984	1.64924	5.21536
2.73	7.4529	1.65227	5.22494
2.74	7.5076	1.65529	5.23450
2.75	7.5625	1.65831	5.24404
2.76	7.6176	1.66132	5.25357
2.77	7.6729	1.66433	5.26308
2.78	7.7284	1.66733	5.27257
2.79	7.7841	1.67033	5.28205
<b>2.80</b>	<b>7.8400</b>	<b>1.67332</b>	<b>5.29150</b>
2.81	7.8961	1.67631	5.30094
2.82	7.9524	1.67929	5.31037
2.83	8.0089	1.68226	5.31977
2.84	8.0656	1.68523	5.32917
2.85	8.1225	1.68819	5.33854
2.86	8.1796	1.69115	5.34790
2.87	8.2369	1.69411	5.35724
2.88	8.2944	1.69706	5.36656
2.89	8.3521	1.70000	5.37587
<b>2.90</b>	<b>8.4100</b>	<b>1.70294</b>	<b>5.38516</b>
2.91	8.4681	1.70587	5.39444
2.92	8.5264	1.70880	5.40370
2.93	8.5849	1.71172	5.41295
2.94	8.6436	1.71464	5.42218
2.95	8.7025	1.71756	5.43139
2.96	8.7616	1.72047	5.44059
2.97	8.8209	1.72337	5.44977
2.98	8.8804	1.72627	5.45894
2.99	8.9401	1.72916	5.46809
<b>3.00</b>	<b>9.0000</b>	<b>1.73205</b>	<b>5.47723</b>
<b>N</b>	<b>N<sup>2</sup></b>	<b>√N</b>	<b>√10N</b>

Table G. SQUARES AND SQUARE ROOTS—(Continued)

N	N <sup>2</sup>	$\sqrt{N}$	$\sqrt{10N}$
3.00	9.0000	1.73205	5.47723
3.01	9.0601	1.73494	5.48655
3.02	9.1204	1.73781	5.49545
3.03	9.1809	1.74069	5.50454
3.04	9.2416	1.74356	5.51362
3.05	9.3025	1.74642	5.52268
3.06	9.3636	1.74929	5.53173
3.07	9.4249	1.75214	5.54076
3.08	9.4864	1.75499	5.54977
3.09	9.5481	1.75784	5.55878
3.10	9.6100	1.76068	5.56776
3.11	9.6721	1.76352	5.57674
3.12	9.7344	1.76635	5.58570
3.13	9.7969	1.76918	5.59464
3.14	9.8596	1.77200	5.60357
3.15	9.9225	1.77482	5.61249
3.16	9.9856	1.77764	5.62139
3.17	10.0489	1.78045	5.63028
3.18	10.1124	1.78326	5.63915
3.19	10.1761	1.78606	5.64801
3.20	10.2400	1.78885	5.65685
3.21	10.3041	1.79165	5.66569
3.22	10.3684	1.79444	5.67450
3.23	10.4329	1.79722	5.68331
3.24	10.4976	1.80000	5.69210
3.25	10.5625	1.80278	5.70088
3.26	10.6276	1.80555	5.70964
3.27	10.6929	1.80831	5.71839
3.28	10.7584	1.81108	5.72713
3.29	10.8241	1.81384	5.73585
3.30	10.8900	1.81659	5.74456
3.31	10.9561	1.81934	5.75326
3.32	11.0224	1.82209	5.76194
3.33	11.0889	1.82483	5.77062
3.34	11.1556	1.82757	5.77927
3.35	11.2225	1.83030	5.78792
3.36	11.2896	1.83303	5.79655
3.37	11.3569	1.83576	5.80517
3.38	11.4244	1.83848	5.81378
3.39	11.4921	1.84120	5.82237
3.40	11.5600	1.84391	5.83095
3.41	11.6281	1.84662	5.83952
3.42	11.6964	1.84932	5.84808
3.43	11.7649	1.85203	5.85662
3.44	11.8336	1.85472	5.86515
3.45	11.9025	1.85742	5.87367
3.46	11.9716	1.86011	5.88218
3.47	12.0409	1.86279	5.89067
3.48	12.1104	1.86548	5.89915
3.49	12.1801	1.86816	5.90762
3.50	12.2500	1.87083	5.91608
N	N <sup>2</sup>	$\sqrt{N}$	$\sqrt{10N}$
3.50	12.2500	1.87083	5.91608
3.51	12.3201	1.87350	5.92453
3.52	12.3904	1.87617	5.93296
3.53	12.4609	1.87883	5.94138
3.54	12.5316	1.88149	5.94979
3.55	12.6025	1.88414	5.95819
3.56	12.6736	1.88680	5.96657
3.57	12.7449	1.88944	5.97495
3.58	12.8164	1.89209	5.98331
3.59	12.8881	1.89473	5.99166
3.60	12.9600	1.89737	6.00000
3.61	13.0321	1.90000	6.00833
3.62	13.1044	1.90263	6.01664
3.63	13.1769	1.90526	6.02495
3.64	13.2496	1.90788	6.03324
3.65	13.3225	1.91050	6.04152
3.66	13.3956	1.91311	6.04979
3.67	13.4689	1.91572	6.05805
3.68	13.5424	1.91833	6.06630
3.69	13.6161	1.92094	6.07454
3.70	13.6900	1.92354	6.08276
3.71	13.7641	1.92614	6.09098
3.72	13.8384	1.92873	6.09918
3.73	13.9129	1.93132	6.10737
3.74	13.9876	1.93391	6.11555
3.75	14.0625	1.93649	6.12372
3.76	14.1376	1.93907	6.13188
3.77	14.2129	1.94165	6.14003
3.78	14.2884	1.94422	6.14817
3.79	14.3641	1.94679	6.15630
3.80	14.4400	1.94936	6.16441
3.81	14.5161	1.95192	6.17252
3.82	14.5924	1.95448	6.18061
3.83	14.6689	1.95704	6.18870
3.84	14.7456	1.95959	6.19677
3.85	14.8225	1.96214	6.20484
3.86	14.8996	1.96469	6.21289
3.87	14.9769	1.96723	6.22093
3.88	15.0544	1.96977	6.22896
3.89	15.1321	1.97231	6.23699
3.90	15.2100	1.97484	6.24500
3.91	15.2881	1.97737	6.25300
3.92	15.3664	1.97990	6.26099
3.93	15.4449	1.98242	6.26897
3.94	15.5236	1.98494	6.27694
3.95	15.6025	1.98746	6.28490
3.96	15.6816	1.98997	6.29285
3.97	15.7609	1.99249	6.30079
3.98	15.8404	1.99499	6.30872
3.99	15.9201	1.99750	6.31664
4.00	16.0000	2.00000	6.32456
N	N <sup>2</sup>	$\sqrt{N}$	$\sqrt{10N}$



Table G. SQUARES AND SQUARE ROOTS—(Continued)

N	N <sup>2</sup>	√N	√10N
4.00	16.0000	2.00000	6.32456
4.01	16.0801	2.00250	6.33246
4.02	16.1604	2.00499	6.34035
4.03	16.2409	2.00749	6.34823
4.04	16.3216	2.00998	6.35610
4.05	16.4025	2.01246	6.36396
4.06	16.4836	2.01494	6.37181
4.07	16.5649	2.01742	6.37966
4.08	16.6464	2.01990	6.38749
4.09	16.7281	2.02237	6.39531
4.10	16.8100	2.02485	6.40312
4.11	16.8921	2.02731	6.41093
4.12	16.9744	2.02978	6.41872
4.13	17.0569	2.03224	6.42651
4.14	17.1396	2.03470	6.43428
4.15	17.2225	2.03715	6.44205
4.16	17.3056	2.03961	6.44981
4.17	17.3889	2.04206	6.45755
4.18	17.4724	2.04450	6.46529
4.19	17.5561	2.04695	6.47302
4.20	17.6400	2.04939	6.48074
4.21	17.7241	2.05183	6.48845
4.22	17.8084	2.05426	6.49615
4.23	17.8929	2.05670	6.50384
4.24	17.9776	2.05913	6.51153
4.25	18.0625	2.06155	6.51920
4.26	18.1476	2.06398	6.52687
4.27	18.2329	2.06640	6.53452
4.28	18.3184	2.06882	6.54217
4.29	18.4041	2.07123	6.54981
4.30	18.4900	2.07364	6.55744
4.31	18.5761	2.07605	6.56506
4.32	18.6624	2.07846	6.57267
4.33	18.7489	2.08087	6.58027
4.34	18.8356	2.08327	6.58787
4.35	18.9225	2.08567	6.59545
4.36	19.0096	2.08806	6.60303
4.37	19.0969	2.09045	6.61060
4.38	19.1844	2.09284	6.61816
4.39	19.2721	2.09523	6.62571
4.40	19.3600	2.09762	6.63325
4.41	19.4481	2.10000	6.64078
4.42	19.5364	2.10238	6.64831
4.43	19.6249	2.10476	6.65582
4.44	19.7136	2.10713	6.66333
4.45	19.8025	2.10950	6.67083
4.46	19.8916	2.11187	6.67832
4.47	19.9809	2.11424	6.68581
4.48	20.0704	2.11660	6.69328
4.49	20.1601	2.11896	6.70075
4.50	20.2500	2.12132	6.70820
N	N <sup>2</sup>	√N	√10N

N	N <sup>2</sup>	√N	√10N
4.50	20.2500	2.12132	6.70820
4.51	20.3401	2.12368	6.71565
4.52	20.4304	2.12603	6.72309
4.53	20.5209	2.12838	6.73053
4.54	20.6116	2.13073	6.73795
4.55	20.7025	2.13307	6.74537
4.56	20.7936	2.13542	6.75278
4.57	20.8849	2.13776	6.76018
4.58	20.9764	2.14009	6.76757
4.59	21.0681	2.14243	6.77495
4.60	21.1600	2.14476	6.78233
4.61	21.2521	2.14709	6.78970
4.62	21.3444	2.14942	6.79706
4.63	21.4369	2.15174	6.80441
4.64	21.5296	2.15407	6.81175
4.65	21.6225	2.15639	6.81909
4.66	21.7156	2.15870	6.82642
4.67	21.8089	2.16102	6.83374
4.68	21.9024	2.16333	6.84105
4.69	21.9961	2.16564	6.84836
4.70	22.0900	2.16795	6.85565
4.71	22.1841	2.17025	6.86294
4.72	22.2784	2.17256	6.87023
4.73	22.3729	2.17486	6.87750
4.74	22.4676	2.17715	6.88477
4.75	22.5625	2.17946	6.89202
4.76	22.6576	2.18174	6.89928
4.77	22.7529	2.18403	6.90652
4.78	22.8484	2.18632	6.91375
4.79	22.9441	2.18861	6.92098
4.80	23.0400	2.19089	6.92820
4.81	23.1361	2.19317	6.93542
4.82	23.2324	2.19545	6.94262
4.83	23.3289	2.19773	6.94982
4.84	23.4256	2.20000	6.95701
4.85	23.5225	2.20227	6.96419
4.86	23.6196	2.20454	6.97137
4.87	23.7169	2.20681	6.97854
4.88	23.8144	2.20907	6.98570
4.89	23.9121	2.21133	6.99285
4.90	24.0100	2.21359	7.00000
4.91	24.1081	2.21585	7.00714
4.92	24.2064	2.21811	7.01427
4.93	24.3049	2.22036	7.02140
4.94	24.4036	2.22261	7.02851
4.95	24.5025	2.22486	7.03562
4.96	24.6016	2.22711	7.04273
4.97	24.7009	2.22935	7.04982
4.98	24.8004	2.23159	7.05691
4.99	24.9001	2.23383	7.06399
5.00	25.0000	2.23607	7.07107
N	N <sup>2</sup>	√N	√10N



Table G. SQUARES AND SQUARE ROOTS—(Continued)

N	N <sup>2</sup>	$\sqrt{N}$	$\sqrt{10N}$	N	N <sup>2</sup>	$\sqrt{N}$	$\sqrt{10N}$
5.00	25.0000	2.23607	7.07107	5.50	30.2500	2.34521	7.41620
5.01	25.1001	2.23830	7.07814	5.51	30.3601	2.34734	7.42294
5.02	25.2004	2.24054	7.08520	5.52	30.4704	2.34947	7.42967
5.03	25.3009	2.24277	7.09225	5.53	30.5809	2.35160	7.43640
5.04	25.4016	2.24499	7.09930	5.54	30.6916	2.35372	7.44312
5.05	25.5025	2.24722	7.10634	5.55	30.8025	2.35584	7.44983
5.06	25.6036	2.24944	7.11337	5.56	30.9136	2.35797	7.45654
5.07	25.7049	2.25167	7.12039	5.57	31.0249	2.36008	7.46324
5.08	25.8064	2.25389	7.12741	5.58	31.1364	2.36220	7.46994
5.09	25.9081	2.25610	7.13442	5.59	31.2481	2.36432	7.47663
5.10	26.0100	2.25832	7.14143	5.60	31.3600	2.36643	7.48331
5.11	26.1121	2.26053	7.14843	5.61	31.4721	2.36854	7.48999
5.12	26.2144	2.26274	7.15542	5.62	31.5844	2.37065	7.49667
5.13	26.3169	2.26495	7.16240	5.63	31.6969	2.37276	7.50333
5.14	26.4196	2.26716	7.16938	5.64	31.8096	2.37487	7.50999
5.15	26.5225	2.26936	7.17635	5.65	31.9225	2.37697	7.51665
5.16	26.6256	2.27156	7.18331	5.66	32.0356	2.37908	7.52330
5.17	26.7289	2.27376	7.19027	5.67	32.1489	2.38118	7.52994
5.18	26.8324	2.27596	7.19722	5.68	32.2624	2.38328	7.53658
5.19	26.9361	2.27816	7.20417	5.69	32.3761	2.38537	7.54321
5.20	27.0400	2.28035	7.21110	5.70	32.4900	2.38747	7.54983
5.21	27.1441	2.28254	7.21803	5.71	32.6041	2.38956	7.55645
5.22	27.2484	2.28473	7.22496	5.72	32.7184	2.39165	7.56307
5.23	27.3529	2.28692	7.23187	5.73	32.8329	2.39374	7.56968
5.24	27.4576	2.28910	7.23878	5.74	32.9476	2.39583	7.57628
5.25	27.5625	2.29129	7.24569	5.75	33.0625	2.39792	7.58288
5.26	27.6676	2.29347	7.25259	5.76	33.1776	2.40000	7.58947
5.27	27.7729	2.29565	7.25948	5.77	33.2929	2.40208	7.59605
5.28	27.8784	2.29783	7.26636	5.78	33.4084	2.40416	7.60263
5.29	27.9841	2.30000	7.27324	5.79	33.5241	2.40624	7.60920
5.30	28.0900	2.30217	7.28011	5.80	33.6400	2.40832	7.61577
5.31	28.1961	2.30434	7.28697	5.81	33.7561	2.41039	7.62234
5.32	28.3024	2.30651	7.29383	5.82	33.8724	2.41247	7.62889
5.33	28.4089	2.30868	7.30068	5.83	33.9889	2.41454	7.63544
5.34	28.5156	2.31084	7.30753	5.84	34.1056	2.41661	7.64199
5.35	28.6225	2.31301	7.31437	5.85	34.2225	2.41868	7.64853
5.36	28.7296	2.31517	7.32120	5.86	34.3396	2.42074	7.65506
5.37	28.8369	2.31733	7.32803	5.87	34.4569	2.42281	7.66159
5.38	28.9444	2.31948	7.33485	5.88	34.5744	2.42487	7.66812
5.39	29.0521	2.32164	7.34166	5.89	34.6921	2.42693	7.67463
5.40	29.1600	2.32379	7.34847	5.90	34.8100	2.42899	7.68115
5.41	29.2681	2.32594	7.35527	5.91	34.9281	2.43105	7.68765
5.42	29.3764	2.32809	7.36206	5.92	35.0464	2.43311	7.69415
5.43	29.4849	2.33024	7.36885	5.93	35.1649	2.43516	7.70065
5.44	29.5936	2.33238	7.37564	5.94	35.2836	2.43721	7.70714
5.45	29.7025	2.33452	7.38241	5.95	35.4025	2.43926	7.71362
5.46	29.8116	2.33666	7.38918	5.96	35.5216	2.44131	7.72010
5.47	29.9209	2.33880	7.39594	5.97	35.6409	2.44336	7.72658
5.48	30.0304	2.34094	7.40270	5.98	35.7604	2.44540	7.73305
5.49	30.1401	2.34307	7.40945	5.99	35.8801	2.44745	7.73951
5.50	30.2500	2.34521	7.41620	6.00	36.0000	2.44949	7.74597
N	N <sup>2</sup>	$\sqrt{N}$	$\sqrt{10N}$	N	N <sup>2</sup>	$\sqrt{N}$	$\sqrt{10N}$

Table G. SQUARES AND SQUARE ROOTS—(Continued)

N	N <sup>2</sup>	√N	√10N
6.00	36.0000	2.44949	7.74597
6.01	36.1201	2.45153	7.75242
6.02	36.2404	2.45357	7.75887
6.03	36.3609	2.45561	7.76531
6.04	36.4816	2.45764	7.77174
6.05	36.6025	2.45967	7.77817
6.06	36.7236	2.46171	7.78460
6.07	36.8449	2.46374	7.79102
6.08	36.9664	2.46577	7.79744
6.09	37.0881	2.46779	7.80385
6.10	37.2100	2.46982	7.81025
6.11	37.3321	2.47184	7.81665
6.12	37.4544	2.47386	7.82304
6.13	37.5769	2.47588	7.82943
6.14	37.6996	2.47790	7.83582
6.15	37.8225	2.47992	7.84219
6.16	37.9456	2.48193	7.84857
6.17	38.0689	2.48395	7.85493
6.18	38.1924	2.48596	7.86130
6.19	38.3161	2.48797	7.86766
6.20	38.4400	2.48998	7.87401
6.21	38.5641	2.49199	7.88036
6.22	38.6884	2.49399	7.88670
6.23	38.8129	2.49600	7.89303
6.24	38.9376	2.49800	7.89937
6.25	39.0625	2.50000	7.90569
6.26	39.1876	2.50200	7.91202
6.27	39.3129	2.50400	7.91833
6.28	39.4384	2.50599	7.92465
6.29	39.5641	2.50799	7.93095
6.30	39.6900	2.50998	7.93725
6.31	39.8161	2.51197	7.94355
6.32	39.9424	2.51396	7.94984
6.33	40.0689	2.51595	7.95613
6.34	40.1956	2.51794	7.96241
6.35	40.3225	2.51992	7.96869
6.36	40.4496	2.52190	7.97496
6.37	40.5769	2.52389	7.98123
6.38	40.7044	2.52587	7.98749
6.39	40.8321	2.52784	7.99375
6.40	40.9600	2.52982	8.00000
6.41	41.0881	2.53180	8.00625
6.42	41.2164	2.53377	8.01249
6.43	41.3449	2.53574	8.01873
6.44	41.4736	2.53772	8.02496
6.45	41.6025	2.53969	8.03119
6.46	41.7316	2.54165	8.03741
6.47	41.8609	2.54362	8.04363
6.48	41.9904	2.54558	8.04984
6.49	42.1201	2.54755	8.05605
6.50	42.2500	2.54951	8.06226
N	N <sup>2</sup>	√N	√10N
6.50	42.2500	2.54951	8.06226
6.51	42.3801	2.55147	8.06846
6.52	42.5104	2.55343	8.07465
6.53	42.6409	2.55539	8.08084
6.54	42.7716	2.55734	8.08703
6.55	42.9025	2.55930	8.09321
6.56	43.0336	2.56125	8.09938
6.57	43.1649	2.56320	8.10555
6.58	43.2964	2.56515	8.11172
6.59	43.4281	2.56710	8.11788
6.60	43.5600	2.56905	8.12404
6.61	43.6921	2.57099	8.13019
6.62	43.8244	2.57294	8.13634
6.63	43.9569	2.57488	8.14248
6.64	44.0896	2.57682	8.14862
6.65	44.2225	2.57876	8.15475
6.66	44.3556	2.58070	8.16088
6.67	44.4889	2.58263	8.16701
6.68	44.6224	2.58457	8.17315
6.69	44.7561	2.58650	8.17924
6.70	44.8900	2.58844	8.18535
6.71	45.0241	2.59037	8.19146
6.72	45.1584	2.59230	8.19756
6.73	45.2929	2.59422	8.20366
6.74	45.4276	2.59615	8.20975
6.75	45.5625	2.59808	8.21584
6.76	45.6976	2.60000	8.22192
6.77	45.8329	2.60192	8.22800
6.78	45.9684	2.60384	8.23408
6.79	46.1041	2.60576	8.24015
6.80	46.2400	2.60768	8.24621
6.81	46.3761	2.60960	8.25227
6.82	46.5124	2.61151	8.25833
6.83	46.6489	2.61343	8.26438
6.84	46.7856	2.61534	8.27043
6.85	46.9225	2.61725	8.27647
6.86	47.0596	2.61916	8.28251
6.87	47.1969	2.62107	8.28855
6.88	47.3344	2.62298	8.29458
6.89	47.4721	2.62488	8.30060
6.90	47.6100	2.62679	8.30662
6.91	47.7481	2.62869	8.31264
6.92	47.8864	2.63059	8.31865
6.93	48.0249	2.63249	8.32466
6.94	48.1636	2.63439	8.33067
6.95	48.3025	2.63629	8.33667
6.96	48.4416	2.63818	8.34266
6.97	48.5809	2.64008	8.34865
6.98	48.7204	2.64197	8.35464
6.99	48.8601	2.64386	8.36062
7.00	49.0000	2.64575	8.36660
N	N <sup>2</sup>	√N	√10N

Table G. SQUARES AND SQUARE ROOTS—(Continued)

N	N <sup>2</sup>	$\sqrt{N}$	$\sqrt{10N}$
<b>7.00</b>	49.0000	2.64575	8.36660
7.01	49.1401	2.64764	8.37257
7.02	49.2804	2.64953	8.37854
7.03	49.4209	2.65141	8.38451
7.04	49.5616	2.65330	8.39047
7.05	49.7025	2.65518	8.39643
7.06	49.8436	2.65707	8.40238
7.07	49.9849	2.65895	8.40833
7.08	50.1264	2.66083	8.41427
7.09	50.2681	2.66271	8.42021
<b>7.10</b>	50.4100	2.66458	8.42615
7.11	50.5521	2.66646	8.43208
7.12	50.6944	2.66833	8.43801
7.13	50.8369	2.67021	8.44393
7.14	50.9796	2.67208	8.44985
7.15	51.1225	2.67395	8.45577
7.16	51.2656	2.67582	8.46168
7.17	51.4089	2.67769	8.46759
7.18	51.5524	2.67956	8.47349
7.19	51.6961	2.68142	8.47939
<b>7.20</b>	51.8400	2.68328	8.48528
7.21	51.9841	2.68514	8.49117
7.22	52.1284	2.68701	8.49706
7.23	52.2729	2.68887	8.50294
7.24	52.4176	2.69072	8.50882
7.25	52.5625	2.69258	8.51469
7.26	52.7076	2.69444	8.52056
7.27	52.8529	2.69629	8.52643
7.28	52.9984	2.69815	8.53229
7.29	53.1441	2.70000	8.53815
<b>7.30</b>	53.2900	2.70185	8.54400
7.31	53.4361	2.70370	8.54985
7.32	53.5824	2.70555	8.55570
7.33	53.7289	2.70740	8.56154
7.34	53.8756	2.70924	8.56738
7.35	54.0225	2.71109	8.57321
7.36	54.1696	2.71293	8.57904
7.37	54.3169	2.71477	8.58487
7.38	54.4644	2.71662	8.59069
7.39	54.6121	2.71846	8.59651
<b>7.40</b>	54.7600	2.72029	8.60233
7.41	54.9081	2.72213	8.60814
7.42	55.0564	2.72397	8.61394
7.43	55.2049	2.72580	8.61974
7.44	55.3536	2.72764	8.62554
7.45	55.5025	2.72947	8.63134
7.46	55.6516	2.73130	8.63713
7.47	55.8009	2.73313	8.64292
7.48	55.9504	2.73496	8.64870
7.49	56.1001	2.73679	8.65448
<b>7.50</b>	56.2500	2.73861	8.66025
N	N <sup>2</sup>	$\sqrt{N}$	$\sqrt{10N}$

N	N <sup>2</sup>	$\sqrt{N}$	$\sqrt{10N}$
<b>7.50</b>	56.2500	2.73861	8.66025
7.51	56.4001	2.74044	8.66603
7.52	56.5504	2.74226	8.67179
7.53	56.7009	2.74408	8.67756
7.54	56.8516	2.74591	8.68332
7.55	57.0025	2.74773	8.68907
7.56	57.1536	2.74955	8.69483
7.57	57.3049	2.75136	8.70057
7.58	57.4564	2.75318	8.70632
7.59	57.6081	2.75500	8.71206
<b>7.60</b>	57.7600	2.75681	8.71780
7.61	57.9121	2.75862	8.72353
7.62	58.0644	2.76043	8.72926
7.63	58.2169	2.76225	8.73499
7.64	58.3696	2.76405	8.74071
7.65	58.5225	2.76586	8.74643
7.66	58.6756	2.76767	8.75214
7.67	58.8289	2.76948	8.75785
7.68	58.9824	2.77128	8.76356
7.69	59.1361	2.77308	8.76926
<b>7.70</b>	59.2900	2.77489	8.77496
7.71	59.4441	2.77669	8.78066
7.72	59.5984	2.77849	8.78635
7.73	59.7529	2.78029	8.79204
7.74	59.9076	2.78209	8.79773
7.75	60.0625	2.78388	8.80341
7.76	60.2176	2.78568	8.80909
7.77	60.3729	2.78747	8.81476
7.78	60.5284	2.78927	8.82043
7.79	60.6841	2.79106	8.82610
<b>7.80</b>	60.8400	2.79285	8.83176
7.81	60.9961	2.79464	8.83742
7.82	61.1524	2.79643	8.84308
7.83	61.3089	2.79821	8.84873
7.84	61.4656	2.80000	8.85438
7.85	61.6225	2.80179	8.86002
7.86	61.7796	2.80357	8.86566
7.87	61.9369	2.80535	8.87130
7.88	62.0944	2.80713	8.87694
7.89	62.2521	2.80891	8.88257
<b>7.90</b>	62.4100	2.81069	8.88819
7.91	62.5681	2.81247	8.89382
7.92	62.7264	2.81425	8.89944
7.93	62.8849	2.81603	8.90505
7.94	63.0436	2.81780	8.91067
7.95	63.2025	2.81957	8.91628
7.96	63.3616	2.82135	8.92188
7.97	63.5209	2.82312	8.92749
7.98	63.6804	2.82489	8.93308
7.99	63.8401	2.82666	8.93868
<b>8.00</b>	64.0000	2.82843	8.94427
N	N <sup>2</sup>	$\sqrt{N}$	$\sqrt{10N}$

Table G. SQUARES AND SQUARE ROOTS—(Continued)

N	N <sup>2</sup>	√N	√10N
8.00	64.0000	2.82843	8.94427
8.01	64.1601	2.83019	8.94986
8.02	64.3204	2.83196	8.95545
8.03	64.4809	2.83373	8.96103
8.04	64.6416	2.83549	8.96660
8.05	64.8025	2.83725	8.97218
8.06	64.9636	2.83901	8.97775
8.07	65.1249	2.84077	8.98332
8.08	65.2864	2.84253	8.98888
8.09	65.4481	2.84429	8.99444
8.10	65.6100	2.84605	9.00000
8.11	65.7721	2.84781	9.00555
8.12	65.9344	2.84956	9.01110
8.13	66.0969	2.85132	9.01665
8.14	66.2596	2.85307	9.02219
8.15	66.4225	2.85482	9.02774
8.16	66.5856	2.85657	9.03327
8.17	66.7489	2.85832	9.03881
8.18	66.9124	2.86007	9.04434
8.19	67.0761	2.86182	9.04986
8.20	67.2400	2.86356	9.05539
8.21	67.4041	2.86531	9.06091
8.22	67.5684	2.86705	9.06642
8.23	67.7329	2.86880	9.07193
8.24	67.8976	2.87054	9.07744
8.25	68.0625	2.87228	9.08295
8.26	68.2276	2.87402	9.08845
8.27	68.3929	2.87576	9.09395
8.28	68.5584	2.87750	9.09945
8.29	68.7241	2.87924	9.10494
8.30	68.8900	2.88097	9.11043
8.31	69.0561	2.88271	9.11592
8.32	69.2224	2.88444	9.12140
8.33	69.3889	2.88617	9.12688
8.34	69.5556	2.88791	9.13236
8.35	69.7225	2.88964	9.13783
8.36	69.8896	2.89137	9.14330
8.37	70.0569	2.89310	9.14877
8.38	70.2244	2.89482	9.15423
8.39	70.3921	2.89655	9.15969
8.40	70.5600	2.89828	9.16515
8.41	70.7281	2.90000	9.17061
8.42	70.8964	2.90172	9.17606
8.43	71.0649	2.90345	9.18150
8.44	71.2336	2.90517	9.18695
8.45	71.4025	2.90689	9.19239
8.46	71.5716	2.90861	9.19783
8.47	71.7409	2.91033	9.20326
8.48	71.9104	2.91204	9.20869
8.49	72.0801	2.91376	9.21412
8.50	72.2500	2.91548	9.21954
N	N <sup>2</sup>	√N	√10N

8.50	72.2500	2.91548	9.21954
8.51	72.4201	2.91719	9.22497
8.52	72.5904	2.91890	9.23038
8.53	72.7609	2.92062	9.23580
8.54	72.9316	2.92233	9.24121
8.55	73.1025	2.92404	9.24662
8.56	73.2736	2.92575	9.25203
8.57	73.4449	2.92746	9.25743
8.58	73.6164	2.92916	9.26283
8.59	73.7881	2.93087	9.26823
8.60	73.9600	2.93258	9.27362
8.61	74.1321	2.93428	9.27901
8.62	74.3044	2.93598	9.28440
8.63	74.4769	2.93769	9.28978
8.64	74.6496	2.93939	9.29516
8.65	74.8225	2.94109	9.30054
8.66	74.9956	2.94279	9.30591
8.67	75.1689	2.94449	9.31128
8.68	75.3424	2.94618	9.31665
8.69	75.5161	2.94788	9.32202
8.70	75.6900	2.94958	9.32738
8.71	75.8641	2.95127	9.33274
8.72	76.0384	2.95296	9.33809
8.73	76.2129	2.95466	9.34345
8.74	76.3876	2.95635	9.34880
8.75	76.5625	2.95804	9.35414
8.76	76.7376	2.95973	9.35949
8.77	76.9129	2.96142	9.36483
8.78	77.0884	2.96311	9.37017
8.79	77.2641	2.96479	9.37550
8.80	77.4400	2.96648	9.38083
8.81	77.6161	2.96816	9.38616
8.82	77.7924	2.96985	9.39149
8.83	77.9689	2.97153	9.39681
8.84	78.1456	2.97321	9.40213
8.85	78.3225	2.97489	9.40744
8.86	78.4996	2.97658	9.41276
8.87	78.6769	2.97825	9.41807
8.88	78.8544	2.97993	9.42338
8.89	79.0321	2.98161	9.42868
8.90	79.2100	2.98329	9.43398
8.91	79.3881	2.98496	9.43928
8.92	79.5664	2.98664	9.44458
8.93	79.7449	2.98831	9.44987
8.94	79.9236	2.98998	9.45516
8.95	80.1025	2.99166	9.46044
8.96	80.2816	2.99333	9.46573
8.97	80.4609	2.99500	9.47101
8.98	80.6404	2.99666	9.47629
8.99	80.8201	2.99833	9.48156
9.00	81.0000	3.00000	9.48683
N	N <sup>2</sup>	√N	√10N



Table G. SQUARES AND SQUARE ROOTS—(Continued)

N	N <sup>2</sup>	√N	√10N
9.00	81.0000	3.00000	9.48683
9.01	81.1801	3.00167	9.49210
9.02	81.3604	3.00333	9.49737
9.03	81.5409	3.00500	9.50263
9.04	81.7216	3.00666	9.50789
9.05	81.9025	3.00832	9.51315
9.06	82.0836	3.00998	9.51840
9.07	82.2649	3.01164	9.52365
9.08	82.4464	3.01330	9.52890
9.09	82.6281	3.01496	9.53415
9.10	82.8100	3.01662	9.53939
9.11	82.9921	3.01828	9.54463
9.12	83.1744	3.01993	9.54987
9.13	83.3569	3.02159	9.55510
9.14	83.5396	3.02324	9.56033
9.15	83.7225	3.02490	9.56556
9.16	83.9056	3.02655	9.57079
9.17	84.0889	3.02820	9.57601
9.18	84.2724	3.02985	9.58123
9.19	84.4561	3.03150	9.58645
9.20	84.6400	3.03315	9.59166
9.21	84.8241	3.03480	9.59687
9.22	85.0084	3.03645	9.60208
9.23	85.1929	3.03809	9.60729
9.24	85.3776	3.03974	9.61249
9.25	85.5625	3.04138	9.61769
9.26	85.7476	3.04302	9.62289
9.27	85.9329	3.04467	9.62808
9.28	86.1184	3.04631	9.63328
9.29	86.3041	3.04795	9.63846
9.30	86.4900	3.04959	9.64365
9.31	86.6761	3.05123	9.64883
9.32	86.8624	3.05287	9.65401
9.33	87.0489	3.05450	9.65919
9.34	87.2356	3.05614	9.66437
9.35	87.4225	3.05778	9.66954
9.36	87.6096	3.05941	9.67471
9.37	87.7969	3.06105	9.67988
9.38	87.9844	3.06268	9.68504
9.39	88.1721	3.06431	9.69020
9.40	88.3600	3.06594	9.69536
9.41	88.5481	3.06757	9.70052
9.42	88.7364	3.06920	9.70567
9.43	88.9249	3.07083	9.71082
9.44	89.1136	3.07246	9.71597
9.45	89.3025	3.07409	9.72111
9.46	89.4916	3.07571	9.72625
9.47	89.6809	3.07734	9.73139
9.48	89.8704	3.07896	9.73653
9.49	90.0601	3.08058	9.74166
9.50	90.2500	3.08221	9.74679
N	N <sup>2</sup>	√N	√10N

N	N <sup>2</sup>	√N	√10N
9.50	90.2500	3.08221	9.74679
9.51	90.4401	3.08383	9.75192
9.52	90.6304	3.08545	9.75705
9.53	90.8209	3.08707	9.76217
9.54	91.0116	3.08869	9.76729
9.55	91.2025	3.09031	9.77241
9.56	91.3936	3.09192	9.77753
9.57	91.5849	3.09354	9.78264
9.58	91.7764	3.09516	9.78775
9.59	91.9681	3.09677	9.79285
9.60	92.1600	3.09839	9.79796
9.61	92.3521	3.10000	9.80306
9.62	92.5444	3.10161	9.80816
9.63	92.7369	3.10322	9.81326
9.64	92.9296	3.10483	9.81835
9.65	93.1225	3.10644	9.82344
9.66	93.3156	3.10805	9.82853
9.67	93.5089	3.10966	9.83362
9.68	93.7024	3.11127	9.83870
9.69	93.8961	3.11288	9.84378
9.70	94.0900	3.11448	9.84886
9.71	94.2841	3.11609	9.85393
9.72	94.4784	3.11769	9.85901
9.73	94.6729	3.11929	9.86408
9.74	94.8676	3.12090	9.86914
9.75	95.0625	3.12250	9.87421
9.76	95.2576	3.12410	9.87927
9.77	95.4529	3.12570	9.88433
9.78	95.6484	3.12730	9.88939
9.79	95.8441	3.12890	9.89444
9.80	96.0400	3.13050	9.89949
9.81	96.2361	3.13209	9.90454
9.82	96.4324	3.13369	9.90959
9.83	96.6289	3.13528	9.91464
9.84	96.8256	3.13688	9.91968
9.85	97.0225	3.13847	9.92472
9.86	97.2196	3.14006	9.92975
9.87	97.4169	3.14166	9.93479
9.88	97.6144	3.14325	9.93982
9.89	97.8121	3.14484	9.94485
9.90	98.0100	3.14643	9.94987
9.91	98.2081	3.14802	9.95490
9.92	98.4064	3.14960	9.95992
9.93	98.6049	3.15119	9.96494
9.94	98.8036	3.15278	9.96995
9.95	99.0025	3.15436	9.97497
9.96	99.2016	3.15595	9.97998
9.97	99.4009	3.15753	9.98499
9.98	99.6004	3.15911	9.98999
9.99	99.8001	3.16070	9.99500
10.00	100.000	3.16228	10.0000
N	N <sup>2</sup>	√N	√10N

# Index

- Alienation, coefficient of, 135
- Analysis of variance, 249-342
  - applications for significance:
    - of correlation, linear, 264-268, 272-275
    - of correlation ratio, 262-264, 272-275
    - of differences:
      - for correlated means, 288-290, 317, 325
      - for independent means, 253-255, 256-258
    - of interaction, 301, 306, 308, 309, 324, 328-335
    - of multiple correlation, 276-279
    - of nonlinearity, 268-275
    - of reliability, 290-294, 310
  - assumptions:
    - homogeneity of variances, 249, 255, 311, 335
    - independent variance estimates, 249, 252
    - normality, 249, 255, 305-306, 328
    - violation of, 255
  - classifications:
    - higher, 337-338
    - one-way or simple, 249-250, 281
    - two-way or double, 281-288
    - three-way or triple, 311-327
  - computation:
    - double classification, 294-296, 298-301
    - groups of unequal size, 261-262
    - simple classification, 256-258
    - single group, 107
    - triple classification, 317-324
  - covariance method, 343-356
    - computation, 350-352
    - and correlation, 345-347
    - degrees of freedom, 345, 350
    - multiple, 354-355
    - regression adjustments, 347-350, 353
- Analysis of variance, covariance method (*Continued*)
  - situations for use, 343-344, 353-354
  - sums of products, 345
  - degrees of freedom, 251-252, 265-266, 269, 276-277, 286-287, 316
  - error term for  $F$ , 293, 303-310, 327-335
  - factorial design, 338
  - interaction, 283, 297
    - higher, 337-338
  - illustrations of, 301-303
    - simple, 297, 301
    - triple, 315
  - Latin square design, 338-342
  - models, 304-305, 327-328
    - components of variance, 304-305, 308-309, 332
    - fixed constants, 304-308, 327-330, 338
    - mixed, 304-305, 309-310, 330-335, 338
  - pooling, 336, 340, 341
  - preliminary tests, 335-337
  - significant  $F$ , meaning of, 253-255
  - sum of squares, 25
    - between-groups, 251
    - breakdown of, 249-252
    - remainder, 286
    - within-groups, 251
  - variance estimates, 94
    - between-groups, 253
    - expected value of, 305
    - interaction, 297
    - meaning of, 252-255
    - remainder, 287
      - as error, 292
    - residual, 267, 287
    - within-cells, 297
    - within-groups, 252-253
- Arbitrary origin, 16-17
- Area sampling, 362
- Arkin, H., 12



- Array, 116, 122  
 Attenuation, 159-160  
 Attributes, 55  
 Average, 1, 16  
 Average deviation, 21  
  
 Bartlett's test, 248  
 Best-fit line, 126-130  
 Beta ( $\beta$ ) coefficients, 172  
 Binomial distribution, 43-46  
   and chi square, 213-214  
   and hypothesis testing, 49-53  
   kurtosis of, 45  
   mean of, 45  
   and normal curve, 46-49  
   and probability, 42-45  
   skewness of, 45  
   standard deviation of, 45  
 Biserial correlation, 192-197  
   assumptions, 194  
   formulas, 193, 196  
   interpretation of, 195  
   and point biserial, 196-197  
   sampling error of, 194, 197  
 Blakeman criterion, 268  
 Brinton, W. C., 12  
 Brown-Spearman formula, 156-157  
  
 Central value (tendency), 14  
   mean, 16-18  
   median, 14-16  
   mode, 14  
 Changes, evaluation of:  
   for categorical data, 56-59  
   by covariance method, 355-356  
   for graduated series, 79-80, 111-112, 355-356  
 Chesire, L., 199n  
 Chi square ( $\chi^2$ ), 212-240  
   additive property of, 226, 230  
   applications as test:  
     of agreement with a priori frequencies, 222-223  
     of changes, 228-230  
     of correlation, 225, 235-236  
     of goodness of fit, 224, 236-238  
     of group differences, 223-224, 225-226, 233-235  
     of independence, 223  
     of several correlated proportions, 232-233  
   assumptions, 221-222  
   and binomial, 213-214  
  
 Chi square ( $\chi^2$ ) (*Continued*)  
   combining of, 226, 230  
   correction to, for continuity, 230-231  
   and critical ratio, 214, 218, 226-228, 229-230  
   degrees of freedom, 216-217, 238-239  
   and discontinuity, 214, 221-222  
   distribution of, 213-221  
     curves, 219  
     mathematical, 218  
   levels of significance, 218-221, 238  
   and normal curve, 214, 221  
   and null hypothesis, 220-221  
   one- vs. two-tailed tests, 231-232  
   and proportions, 226-228  
   table of, 386-387  
 Classification, 5-6; *see also under*  
   Analysis of variance  
 Cochran, W. G., 113, 232n  
 Coded scores, 18, 23  
 Colton, R. R., 12  
 Combined groups:  
   mean for, 19  
   standard deviation for, 26  
 Common elements and correlation, 140-141  
 Comparison of groups, 82-83; *see also*  
   Significance, of differences  
 Confidence coefficient, 98  
 Confidence interval, 95-99, 108, 109, 110-111  
 Confidence level, 98  
 Confidence limits, 95-99, 107-108, 109, 110-111  
 Confounded, 329  
 Contingency coefficient, 203-206  
   and chi square, 203, 224-228, 235-236  
   corrections to, 205-206  
   sampling error of, 206  
   upper limits of, 205  
 Contingency table, 204, 224, 235  
 Continuity, correction for, 48, 52, 54, 58, 61, 230-231  
 Continuous series, 5  
 Correction:  
   for attenuation, 159-160  
   to contingency coefficient, 205-206  
   for continuity, 48, 52, 54, 58, 61, 230-231  
   for grouping, 25  
   for uncontrolled variable, 344

- Correlation and causation, 140, 166-167, 187
- Correlation between:
- categorized variables, 197-206
  - dichotomized and graduated variables, 192-197
  - dichotomized variables, 197-206
  - graduated variables, 118, 207
  - indexes, 161-163
  - means, 86, 88
  - point variables, 202-203
  - standard deviations, 88
- Correlation:
- factors affecting, 144-168
    - errors of measurement, 159-160
    - heterogeneity, 149-150
      - third variable, 164-167
    - indexes, 161-163
    - part-whole, 164
    - range of talent, 149-150
    - sampling errors, 145-147, 264-268
    - selection, 144
  - measures of:
    - biserial, 192-197
    - contingency, 203-206
    - correlation ratio ( $\eta$ ), 207-208, 262-264, 272-275
    - fourfold point, 202-203
    - intraclass, 280
    - multiple, 169-190; *see also* Multiple correlation
    - partial, 164-167
    - point biserial, 196-197
    - product moment, 115-143; *see also* Product moment correlation
    - rank, 208-210
    - tetrachoric, 197-202
- Correlation ratio ( $\eta$ ), 207-208
- computation, 272-275
  - sampling significance of, 262-264, 272-275
- Correlations, averaging of, 148-149
- Covariance, 344; *see also under* Analysis of variance
- Cox, G. M., 113
- Crespi, L., 235
- Critical ratio ( $CR$ ), 54, 58
- and chi square, 214, 218, 226-228, 229-230
  - and  $F$ , 261
  - and  $t$ , 109
- Critical region, 67
- Cumulative frequency distribution, 10
- Curvilinearity, test of, 268-275
- Decile, 20
- Degrees of freedom:
- in analysis of variance, 251-252, 265-266, 269, 276-277, 286-289, 316
  - for chi square, 216-217, 238-240
  - for  $F$ , 245
  - for  $t$  test:
    - for means, 106-107, 111
    - for  $r$ , 146, 167
  - for variance estimate, 106-107
- Deming, W. E., 363n
- Descriptive statistics, 2, 13
- Deviation score, 21
- Differences, *see* Significance, of differences
- Discontinuity, *see* Continuity
- Discrete series, 5
- Discriminant function, 210-211
- Distribution:
- binomial, 43-46
  - chi square, 217-220
  - cumulative, 10
  - expected, 40
  - $F$ , 245
  - frequency, 6
  - mathematical, 40
  - normal, 32-37
  - observed, 40
  - population, 40
  - sampling, 53
  - $t$ , 105-106
  - theoretical, 40
- Distribution-free methods, 357-360
- chi square as, 357
  - for correlated sets, 358-359
  - Mann-Whitney  $U$  test, 359-360
  - "median" test, 358
  - sign test, 357
- Doolittle method, 182-185
- Edwards, A. L., 337n
- Elderton's table of chi square, 221
- Error:
- absolute, 151
  - constant, 151
  - in drawing conclusions, 64-70
  - of estimate, 131-136, 174-176
  - of measurement, 151-154
  - reduction, 88-90, 361-366

Error (*Continued*)

- relative, 151
- sampling, *see under* Standard error
- standard, *see* Standard error
- type I and type II, 65-69
- variable, 150, 154-155

Estimate, error of, 131-133, 174-176

## Estimation:

- interval, 95-99
- point, 94

## Estimator:

- consistency, 94
- efficiency, 94
- unbiased, 94

Eta ( $\eta$ ), 207-208

- computation, 272-275
- sampling significance of, 262-264, 272-275

## Experimental and control data, treatment of:

- matched distributions, 364-365
- own control, 86, 89, 108-109, 288-290, 317, 325
- paired (or matched) cases, 59-60, 86, 89-90, 108-109, 288-290, 317, 325
- randomly drawn, 87, 109-111, 256-259
- sibs and littermates, 86, 108-109, 288-290, 317, 325

Ezekiel, M., 188

 $F$ , or variance ratio, 245

- and critical ratio ( $CR$ ), 261
- degrees of freedom, 245
- distribution, 245
- error term for, 303-310, 327-335
- for group variances, 245-247
- of independent estimates, 249, 252-253
- and  $t$ , 260, 268, 289
- table of, 389-391

## Factorial design, 338

## Fiducial limits, 98

## Finite universe, 99-100

Fisher, R. A., 64, 147, 244, 356, 385-391

## Fitting of line, 126-130

Form vs. form reliability, 157-158, 290-294, 310

## Fourfold point correlation, 202-203

## Fourfold table, 56-57, 198

and changes, 57, 228-230

Fourfold table (*Continued*)

- chi square for, 206, 224
- and contingency, 203, 206
- exact probability for, 241-242
- and point correlation, 202, 206
- and tetrachoric  $r$ , 197-202

## Frequency, 6

- as area, 9-10
- comparison, *see* Chi square
- cumulative, 10
- curve, 8
- distribution, 6
- polygon, 8
- table, 6

## Goodness of fit, 224, 236-238

## Graduated series, 5

## Graphic presentation, 7-12

- histogram, 7
- line graph, 11-12
- ogive, 10
- polygon, 8

## Grouping, 5-6

- and coding, 18
- correction for, 25

## Guessed average, 18

## Heterogeneity and correlation, 149-150, 159, 164-167, 347

## Histogram, 7

## Homoscedasticity, 131, 248

test of, 248

## Horst, P., 337n

## Hypotheses, 50, 61

- alternate, 61-62
- null, 56, 61-62
- one- vs. two-tailed, 62-64, 112, 231-232, 246-247
- research, 61
- statistical, 61

## Hypothesis testing, 49

- by binomial, 49-53
- by chi square, 220-221
- by  $F$  distribution, 245-247, 255
- by normal distribution, 52, 58, 78
- by  $t$  distribution, 106-114

## Independence, test of, 223

## Indexes:

- correlation of, 161-163
- mean of, 162
- standard deviation of, 162

## Inference, statistical, 2, 50, 94

- Interaction, 283, 297, 301-303, 315, 337-338  
     and group profiles, 335
- Intervals:  
     grouping, 5-6  
     limits of, 6-7  
     midpoints of, 7  
     size of, 6
- Intraclass correlation, 280
- Kelley, T. L., 181, 205
- Kendall, M. G., 210
- Kurtosis, 13, 27-30
- Latin square design, 338-342
- Level:  
     of confidence, 98-99  
     of significance, 51, 64-70, 99, 113-114
- Line graph, 11-12
- Linearity of regression, 126  
     test for, 268-275
- McCall, W. A., 39
- McNemar, Q., 379n
- Mann-Whitney *U* test, 359-360
- Matched groups by means of:  
     matched distributions, 364-365  
     paired cases, 59-60, 86, 89, 364, 365  
     randomization, 364-365  
     siblings and twins, 86, 364, 366
- Mean, 16  
     for combined groups, 19  
     computation, 17-18  
     sampling error of, 76-78, 104
- Mean difference, significance of, 79-80, 83-85, 108-109
- Measurement error, 150-154, 200-204
- Median, 14-16
- "Median" test, 358
- Mentzer, E. G., 304n
- Midpoint of interval, 7, 14
- Mode, 14
- Models in analysis of variance, 304-306, 327-328
- Moments, 27-28
- Mood, A. M., 359
- Moses, L., 360
- Moving averages, 8
- Mueller, C. G., 357n
- Multiple correlation, 175-176  
     in covariance, 354-355  
     and determinants, 180-181
- Multiple correlation (*Continued*)  
     and diminishing returns, 188  
     and discriminant function, 211  
     Doolittle method, 181-185  
     error of estimate, 174-176  
     interpretation of, 176  
     limitations, 186-188  
     notation, 189-190  
     numerical solution, 181-185  
     regression equations, 172-174, 180  
     relative weights in, 176-177  
     sampling error of, 185-186, 276-279  
     selection fallacy, 187  
     and shrinkage, 186, 279  
     and suppressant variable, 188-189
- Nonlinearity, test of, 268-275
- Nonparametric methods, 357-360;  
     *see also under* Distribution-free
- Normal correlation, 141-142
- Normal distribution curve, 33  
     area under, 34-36  
     equations for, 33, 35  
     and probability, 46-49  
     table of, 382-383  
     unit form of, 35
- Null hypothesis, 56, 61-62
- Ogive, 10
- One- vs. two-tailed tests, 62-64, 112  
     binomial, 52, 60, 241  
     chi square, 231-232  
     *F* ratio, 246-247
- Paired cases, 86, 89, 364, 365
- Parameter, 2
- Partial correlation, 165-167  
     sampling error of, 167
- Part-whole correlation, 164
- Paterson, D. G., 182n
- Paull, A. E., 336
- Pearson, K., 32, 142n, 198n, 221n
- Percentage, *see* Proportion
- Percentile, 20-21
- Peters, C. C., 112
- Point biserial correlation, 190-197
- Point series, 46
- Power of a test, 68-69
- Prediction, error of, 131-136, 174-176
- Probability, 42  
     addition theorem, 42  
     approximations to, 46-49, 240-241  
     as area, 48-49  
     and binomial, 43-46

- Probability (*Continued*)  
   exact, 240-242  
   and hypothesis testing, 49-53  
   as level of significance, 51  
   multiplication theorem, 42  
   of type I error, 65, 67  
   of type II error, 67-70  
 Probable error, 102-103  
 Product moment correlation, 118  
   assumptions, 126, 131, 136-137,  
     139-140, 141, 143  
   computation, 119-121  
   direction of, 134-135  
   interpretations, in terms of:  
     common elements, 140-141  
     error of estimate, 131-136  
     normal surface, 141-142  
     rate of change, 130  
     variance explained, 137-140  
   limits for, 142, 160-161, 167  
   and prediction, 126-127, 130  
   and regression, 130  
   sampling error of, 145-147, 264-268  
   scatter diagram, 116-118, 122-126  
 Profiles and interaction, 335  
 Proportion, sampling error of, 53-54  
 Proportions as means, 101-102  
  
 Quartile, 20  
 Quartile deviation, 19-20  
 Quota sampling, 363  
  
 Random sampling, 55, 75, 361-362  
 Randomization, 364-365  
 Range, 6, 19  
 Rank correlation, 208-210  
   Kendall's tau, 210  
   Spearman's rho, 208  
   significance of, 210  
 Regression, 130  
   coefficients, 130, 176  
   equations, 129-130, 172, 174, 180  
   test of linearity of, 268-275  
 Reliability, 151-159, 290-294  
   and attenuation, 159-160  
   coefficient of, 151  
   of difference scores, 154  
   error of measurement, 153, 294  
   form vs. form, 157-158  
   range, effect of, 159  
   significance of, 290-294, 310  
   split-half, 156-157  
   test-retest, 156  
  
 Renshaw, M. J., 299  
 Replication, 304-305  
 Residuals, 138, 175, 267, 287, 339  
 Rider, P. R., 304n  
  
 Saffir, M., 199n  
 Sampling, 55  
   distribution, 53, 76  
     binomial as, 53  
     of chi square, 213-221  
     empirical demonstration of, 73-  
       75  
     of  $F$ , 245  
     of  $t$ , 105-106  
   empirical demonstration, 73-75  
   errors, reduction of, 88-90, 361-  
     366  
   for experimental and control  
     groups, 89-90, 363-366  
   from finite universe, 99-100  
   independence of units, 99, 222  
   size of sample required, 70, 90, 113-  
     114  
   from skewed universe, 100-101  
   small samples, 58, 71, 104  
   successive, 76  
   techniques, 361-363  
     area, 363  
     quota, 363  
     random, 55, 75, 361-362  
     stratified, 362-363  
     systematic, 362  
   theory, 55, 75-78  
   variance, 76  
 Scatter diagram, 116-118, 122-126  
 Sheppard's correction, 25  
 Shrinkage of multiple  $r$ , 186, 279  
 Sign test, 357  
 Significance, 51  
   choice of level, 64-70  
   of correlation, 145-147, 264-268  
   of correlation ratio, 262-264, 272-  
     275  
   of differences:  
     for changes, 90-94  
     for correlations, 148  
   for means:  
     correlated, 85, 108-109, 288-  
       290, 317, 325  
     independent, 87, 109-110, 252-  
       262  
     sub- vs. total group, 100

Significance, of differences (*Continued*)

- for proportions:
  - correlated, 56-60, 228-230, 232-233
  - independent, 60-61
- for scores, 154
- for standard deviations, 88, 243-248

## for variances:

- Bartlett's test, 247-248
- correlated, 243-244
- independent, 244-248
- and erroneous conclusions, 65-70
- of interaction, 301, 306, 308, 309, 324, 328-335
- levels, 51, 64-70
- of multiple  $r$ , 185-186, 276-279
- of nonlinearity, 268-275
- of reliability, 290-294, 310

## Skewness, 13, 27-30

- of binomial distribution, 45
- causes of, 30-31
- of sampling distributions:
  - of correlations, 146-147
  - of proportions (or percentages), 54
  - of standard deviations, 105

## Small sample treatment, 58, 71, 104

- of correlation, 146
- of difference:
  - for correlated means, 108-109
  - for independent means, 109-110
  - for variances, 243-248
- of single mean, 107-108

*see also* Analysis of variance

## Smoothing, 8

## Snedecor, G. W., 245

## Spearman-Brown formula, 156-157

## Split-half reliability, 156-157

## Spurious correlation, 163, 164

## Squares and square roots, 392-400

## Standard deviation, 21

- for combined groups, 26
- computation, 22-25
- sampling error of, 81, 243
- Sheppard's correction, 25

## Standard error, 53, 76

- of average deviation, 81
- of correlation measures:
  - biserial, 194
  - multiple, 185
  - product moment, 145
  - tetrachoric, 200
  - $z$  (transformed  $r$ ), 147

Standard error (*Continued*)

- of kurtosis, 82
- of mean, 77, 81
  - from finite universe, 100
  - for stratified sample, 363
- of mean difference, 79, 85
- of median, 81
- of proportion, 53-54
  - from finite universe, 100
  - for stratified sample, 362
- of quartile deviation, 81
- of skewness, 82
- of standard deviation, 81
- of  $z$  (transformed  $r$ ), 147

## Standard error of difference, 59, 83

- for changes, 92
- for means:
  - correlated, 85, 364-365
  - independent, 87
  - sub- vs. total group, 100

## for medians, 88

## for proportions:

- correlated, 56-60
- independent, 60-61

## for scores, 154

## for standard deviations, 88

for  $z$ 's (transformed  $r$ 's), 148

## Standard error of estimate, 131-136, 174-176

## Standard error of measurement, 150-154, 290-294

## Standard score, 34, 37-39

and  $T$  score, 39

## Statistic, 2

## Stratified sampling, 362-363

## "Student," 366

## Successive sampling, 76

Sum of squares, 25, 107; *see also under* Analysis of variance

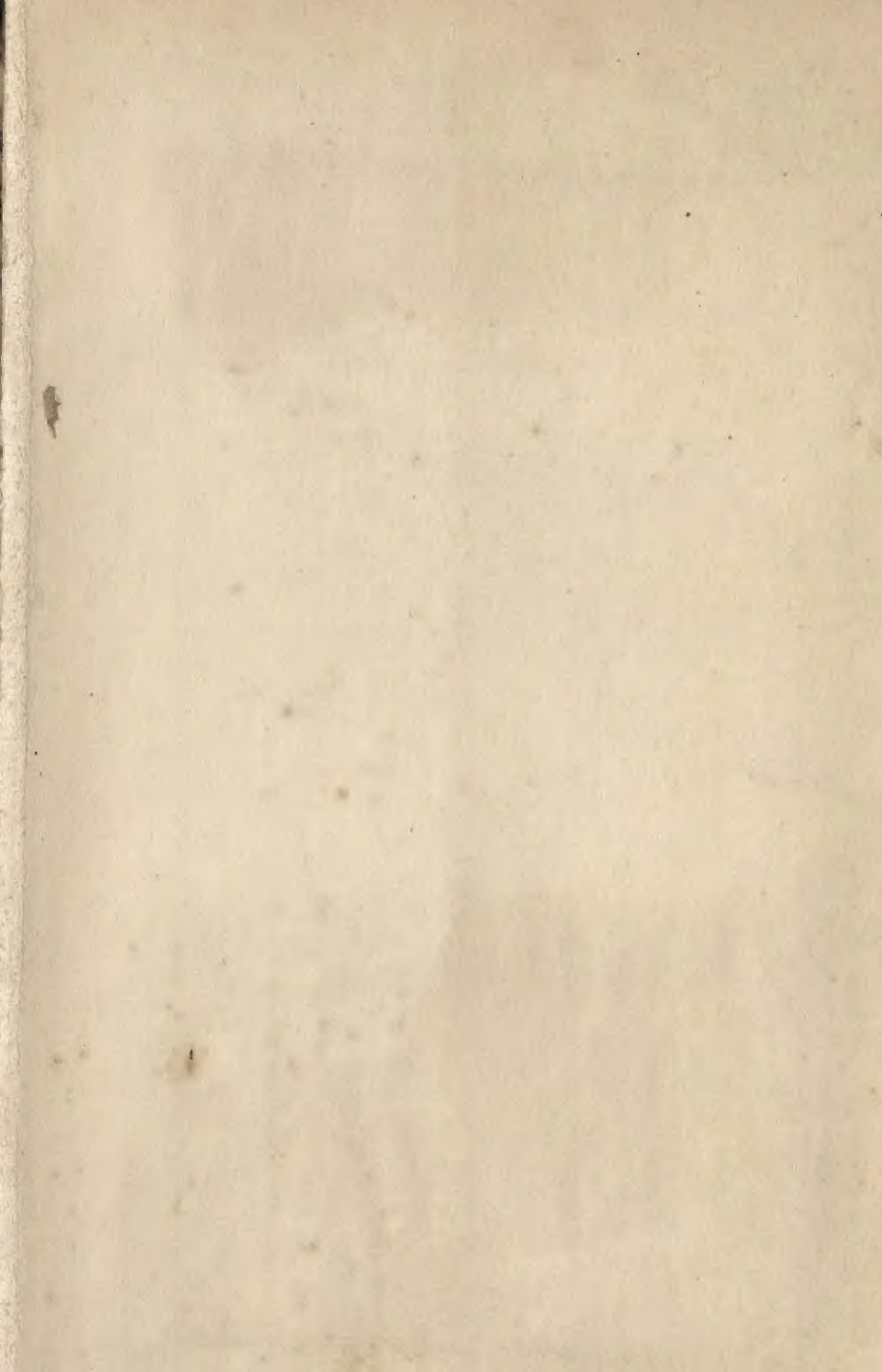
## Suppressant variable, 188-189

 $t$  ratio, 105

- assumptions and limitations in use of, 112-114
- and confidence limits, 107-108, 110
- for correlation, 146
- and critical ratio ( $CR$ ), 109
- degrees of freedom, 106-107, 111
- for difference:
  - in correlated correlations, 148
  - in correlated means, 108-109
  - in correlated variances, 243-244
  - in independent means, 109-110



- t* ratio (*Continued*)  
 distribution of, 105-106  
 and *F*, 260, 268, 289  
 for rank correlation, 210  
 for single mean, 107-108  
 table of, 388
- T* score, 39
- Tabulation, 5-6
- Taubman, R. E., 321n
- Test-retest reliability, 156
- Tetrachoric correlation, 197-202  
 computing diagrams for, 199  
 formula, 199  
 sampling error of, 200-201
- Thorndike, R. L., 355n
- Thurstone, L. L., 199n
- Transformation:  
 mathematical, 191, 357  
 standard scores, 37-38  
*T* scaling, 39
- True score, 151
- Two-tailed tests, 62-64, 112
- U* test, 359-360
- Van Voorhis, W. R., 112
- Variance, 22  
 additive nature of, 137-138  
 and chi square, 243
- Variance (*Continued*)  
 computation, 22-25  
 and correlation, 137-140, 176  
 of difference, 137-138  
 difference between, 243-248  
 estimate of, 94  
 homogeneity of, 248  
 ratio, *see F*  
 sampling of, 243  
 of sum, 137-138  
 theorem, 137-138  
*See also* Analysis of variance
- Variation, 13, 19  
 average deviation, 21  
 coefficient of, 161  
 quartile deviation, 19-20  
 standard deviation, 21
- Walker, E. L., 295, 301
- Wright, Suzanne T., 256, 281
- Yates, F., 363n, 385-391
- Yates's correction for continuity, 230-231
- z*, for difference between standard deviations, 244
- z* score, 34, 37-39
- z* transformation for *r*, 147  
 tables of, 384, 385



Form No. 3.

PSY, RES.L-1

**Bureau of Educational & Psychological  
Research Library.**

The book is to be returned within  
the date stamped last.

29.561  
- 8 JUN 1967

9.2.67

13 APR 1967

311  
MCN  
Form No. 4

BOOK CARD

Coll. No. 311/MCN Accn. No. 805  
Author. McMeman, Ann.  
Title. Psychological Statistics

Date.	Issued to	Returned on
25.6.1957	5660	21.10.57
18 JUN 1957	0257	168

311  
MCN

805



